

Latent Insights into Measurement of Academic Challenges:

An Examination of Perceived Academic Challenges Scale

by

Ramin Rostampour

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Educational Psychology and Leadership Studies

University of Victoria

© Ramin Rostampour, 2024

University of Victoria

All rights reserved. This Dissertation may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.

We acknowledge and respect the Ləkʷəŋən (Songhees and Esquimalt) Peoples on whose territory the university stands, and the Ləkʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Supervisory Committee

Latent Insights into Measurement of Academic Challenges:

An Examination of Perceived Academic Challenges Scale

by

Ramin Rostampour

Supervisory Committee

Dr. Allyson F. Hadwin

Department of Educational Psychology and Leadership Studies, University of Victoria

Supervisor

Dr. Stuart MacDonald

Department of Psychology, University of Victoria

Outside Member

Dr. Phil Winne

Faculty of Education, Simon Fraser University

Outside Member

Abstract

Background: Students often face diverse challenges throughout their educational journey that can significantly impact their learning process. While extensive research has explored self-regulatory perceptions and behaviors that promote student success, there remains a notable gap in understanding and measuring the specific challenges students encounter during their studying.

Objective: To conduct a comprehensive psychometric examination of the *Perceived Academic Challenges* scale, a self-assessment and diagnostic tool which measures six dimensions of perceived academic challenges: motivation, initiating and sustaining engagement, goal and time management, cognition, metacognition, and social and emotional challenges. These dimensions correspond to key self-regulatory areas linked to success in self-regulated learning.

Methods: Four university student samples from six Canadian institutions (total N= 3293) were used. The psychometric process employed both item-level and scale-level analyses, including polytomous item response theory (IRT) and advanced structural equation modeling techniques. Multidimensionality was explored using bifactor and second-order measurement models, and Exploratory Structural Equation Modeling (ESEM). A complementary person-centered study examined intraindividual profiles of challenges. The scale's predictive utility was examined through associations with criterion constructs: students' GPA expectations, self-efficacy for GPA, and actual end-of-semester GPA.

Results: IRT analyses confirmed the precision of individual items in assessing various levels of the intended constructs and flagged imprecise items for potential revision. A bifactor ESEM approach demonstrated the best fit to the data, revealing a general factor underlying responses, with cross-loadings enhancing construct interpretability. The general factor and metacognitive factor were found to be highly overlapping, leading to modifications on the measurement model to anchor the general factor to metacognitive challenges. Findings confirmed the measure's adequate psychometric properties, the adequacy of its total score, and its relevance to the criterion constructs.

Conclusion: The *Academic Challenges* scale was found to represent constructs as intended and to be reliable for assessing students' academic challenges across the six dimensions. This dissertation underscores the critical need to refine self-reported measures of students' experiences, emphasizing their unique insights into subjective experiences that other types of data cannot capture. Recommendations for further improvements of the scale are provided, emphasizing the need for continuous refinement of measures in educational research.

Table of Contents

Supervisory Committee.....	ii
Abstract	iii
Table of Contents.....	iv
List of Tables.....	ix
List of Figures	x
List of Original Manuscripts	xi
Acknowledgements	xii
Dedication.....	xiii
Chapter 1: Theoretical Framework	14
Self regulate Learning (SRL)	14
Winne and Hadwin’s Model of SRL	15
Student Success from the Lens of SRL	19
The Benefit of Self-Report Data for Understanding SRL.....	22
Academic Challenges.....	27
Defining Academic Challenges as a Concept, Construct, and Operationalized Construct.....	28
Conceptual Definition.....	28
Construct Definition	29
Temporal and Contextual Nature of Academic Challenges.....	31
Within Person Dynamics of Academic Challenges	31
Situation Specificity of Academic Challenges	32
Academic Challenges Scale.....	33

Categorization of Academic Challenges.....	34
Importance of Measuring Academic Challenges.....	36
Identifying Gaps in How Measurement Models for the Academic Challenges Have been Operationalized	38
Need for Mapping General and Specific Components in Academic Challenges	41
Need for Capturing Item Interdependencies in Measurement Model of Academic Challenges	44
Need for Validating Arithmetic-Based Scale Scores for Practical Use	45
Need For an Item-Level Analysis of Academic Challenges Scale	46
Need for Assessing Measurement Precision Across Different Levels of Academic Challenges.....	47
Summary of the Chapter.....	48
Chapter 2: Methodological Considerations.....	49
Exploring the Multidimensionality of the Academic Challenges Scale	51
Common Approaches to Construct Validation	51
Exploratory and Confirmatory Factor Analysis.....	52
Modeling a Latent Construct of General Academic Challenge	54
Higher-order Measurement Model and Bifactor Measurement Model	56
Reliability and Dimensionality of Bifactor Models.....	61
Symmetric Bifactor Model Versus S-1 Bifactor Model	63
Integrating Inter-Item Correlational Structures Within the Measurement Model.....	66
Sequential Strategy for Testing Alternative Measurement Models	71
Evaluating Criterion Utility of the Measurement Model	74
Zooming In: Moving from Scale Multidimensional Structure to Item-Level Precision.....	76
Item Response Theory (IRT).....	77
Graded Response Model	78

Graphs Depicting IRT Parameters.....	80
Assumptions of IRT.....	83
Unidimensional versus Multidimensional Item-level Analyses	86
Evaluating the Adequacy of Composite Scores for the Scale Practical Use	87
Evaluating the Psychometric Adequacy of Each Subscale as a Discrete Scale for Research Applications	88
Evaluating the Psychometric Adequacy of the Overall Scale as a Unidimensional Scale	89
The Added Value of Conducting an Item-Level Analysis in Assessment of Measurement Quality	91
Examining Item Performance at Different Levels of the Latent Attribute	91
Assessing Response Scale Effectiveness	92
Developing a Shorter Version of the Scales.....	93
Evaluating Differential Item Functioning.....	93
Sample/Parameter Invariance in IRT and Factor Analysis	94
Computerize Adaptive Testing.....	97
Visual Representation of Measurement Quality	98
Triangulating How Missingness is Handled	98
From Variables to Individuals: Complementary Person-Centered Validation of Academic Challenges Scale	99
Person-centred versus Variable Centered Approaches	99
Benefits of Conducting Complementary Person-Centered Validation of Academic Challenges Scale..	101
Exploring Academic Challenges' Multidimensionality Through Person-Centered Analysis	102
Summary of the Chapter.....	105
Chapter 3: Research Purpose, Context, and Manuscript Overview	107
Research Purpose	107
Research Context and Ethics	109

Pandemic Context.....	110
Participants	110
Measures	115
Overview of Manuscripts	123
Chapter 4: Discussion.....	126
Measurement Quality Versus Psychometric Adequacy	126
Theoretical Considerations	131
Challenges as Psychological Situations	131
Academic Challenges from Reality to the Measured Construct	132
Pragmatic Considerations	134
Organizing and Tailoring Support Based on Academic Challenges	136
Summary of Findings.....	139
Findings from Scale-Level Psychometric Examinations (Studies 1 & 3).....	139
Theory-Driven vs. Data-Driven General Factors	141
Dominance of the General Factor	144
Examining the Criterion Utility of the Measurement Models	144
Findings from the Complementary Person-centered Analysis (Study 3).....	147
Findings from Item-Level Psychometric Examinations (Study 2)	148
Adequacy of Composite Scores for Practical Use of the Measure	151
Criterion Utility of Academic Challenges as Unidimensional Constructs	152
Suggestions For Improving the Scale.....	153
Suggestions for Improving the Measure based on Item-Level Analyses.....	153
General Suggestions for Improving the Measure Based on IRT Findings	153

Suggestions for Refining the Motivation Challenges Scale (MOT) Based on IRT Analysis	156
Suggestions for Refining the Initiating and Sustaining Engagement Scale (ISE) Based on IRT Analysis	158
Suggestions for Refining the Goal and Time Management Challenges Scale (GTM) Based on IRT Analysis ..	160
Suggestions for Refining the Cognitive Challenges Scale (COG) Based on IRT Analysis.....	163
Suggestions for Refining the Metacognitive Challenges Scale (MET) Based on IRT Analysis	165
Suggestions for Refining the Social and Emotional Challenges Scale (SE) Based on IRT Analysis	167
Suggestions for Refining the Unidimensional General Academic Challenges Scale (AC-g) Based on IRT	170
Suggestions for Improving the Scale based on the Scale-Level Psychometric Examinations.....	174
Rethinking Granularity of the Scale	174
Prioritizing Ecological Validity in Defining the General Factor	175
Restructuring the Measurement Model by Integrating Volitional Processes.....	178
Collecting Data on Potential Temporal, Contextual, and Individual Moderators	179
Practical Recommendations for Using the Current Version of the Scale	181
Computing Manifest Scores for the Bifactor Structure of the Scale	182
Future Directions.....	184
Limitations	194
Conclusion.....	200
References	202
Appendix A: Study 1	225
Appendix B: Study 2	299
Appendix C: Study 3	465

List of Tables

Table 1. Definition of the Concepts and Constructs	30
Table 2. Breakdown of Challenges in Each Regulatory Domain	35
Table 3. Potential Cross-Loadings for an Item from Cognitive Challenges	45
Table 4. Key Objectives for Improving Construct Measurement in Academic Challenges Scale.....	49
Table 5. Advantages and Disadvantages of Analytic Methods Discussed.....	74
Table 6. Main Objectives and Research Questions of the Dissertations Studies	108
Table 7. The Psychometric Validity and Statistical Error Across the Analytic Methods Used	130
Table 8. Factors Influencing Tailoring Support Based on Academic Challenges	138
Table 9. Summary of the Findings from Psychometric Examination of the Multidimensionality of the Measure	146
Table 10. Summary of proposed modifications to the Academic Challenges scale based on IRT analysis findings	155
Table 11. Summary of Suggestions for Improving Motivation Challenges Scale Based on IRT Findings	157
Table 12. Suggestions for Improving ISE Items Based on IRT Findings.....	160
Table 13. Suggestions for Improving GTM Items Based on IRT Findings	162
Table 14. Suggestions for Improving COG Items Based on IRT Findings	165
Table 15. Suggestions for Improving MET Items Based on IRT Findings	167
Table 16. Suggestions for Improving SE Items Based on IRT Findings	170
Table 17. AC-g Items Categorized in terms of IRT Information.....	172
Table 18. Suggestions for Improving AC-g Items Based on IRT Findings	174
Table 19. Summary of Suggestions for Improving the Scale based on Scale-Level Analyses	180
Table 20. Summary of Practical Recommendations for Using the Current Version of the Scale.....	181
Table 21. Summary of Future Research Directions	194

List of Figures

Figure 1. The Winne and Hadwin's model of Self-Regulated Learning (1998).....	17
Figure 2. A Conceptual Model Showing the Temporal dynamic of Academic Challenges	31
Figure 3. Bidirectional Process of Construct Operationalization	40
Figure 4. Correlated Factor CFA for the Academic Challenges Scale	53
Figure 5. Higher order (a) and Bifactor (b) Measurement Models for the Academic Challenges.....	56
Figure 6. Proportionality Constraint in Higher-Order Models	58
Figure 7. Disaggregation of General and Specific Factors across Higher-order and Bifactor Models.....	60
Figure 8. An example of Bifactor S-1 Measurement Model with Metacognitive Challenges as the Reference domain	65
Figure 9. Correlated, Second order, and Bifactor CFA and ESEM Models for the Academic Challenges Scale.	73
Figure 10. Boundary Characteristic Curves (BCCs) for an item of the Motivation Challenges Scale	80
Figure 11. Category Characteristic Curves (CCCs) for an item of the Motivation Challenges Scale	81
Figure 12. Item Information Function Graphs (a) and Test Information Curve and Standard Error Graphs (b) ..	81
Figure 13. Conditional Reliability Graph	82
Figure 14. Expected Total Score Graph	83
Figure 15. Monotonicity Plot for an Exemplar Item	85
Figure 16. An Example of Item Expected Score Plot for An Item with A Non-Uniform DIF	94
Figure 17. Person-Centered Mixture Models for the Academic Challenges Scale.....	103
Figure 18. An example of Extracted Profiles Based on First-order and Bifactor Indicators	104
Figure 19. Measurement: Linking Reality to Scores	126
Figure 20. An Example of Students' Academic Challenges Profiles	137
Figure 21. Symmetrical Bifactor ESEM Vs. S-1 Bifactor Model	141
Figure 22. Thresholds Range Plot for AC-g Scale	171
Figure 23. How to Interpret an Individual's Scores on a Specific factor in Relation to their Overall Score	183

List of Original Manuscripts

This dissertation comprises three manuscripts as detailed below:

Manuscript 1: Hadwin, A. F., Rostampour, R., Winne, P. H. (2024). Advancing Self-reports of Self-regulated Learning: Validating New Measures to Assess Students' Beliefs, Practices, and Challenges. Manuscript submitted for publication.

Manuscript 2: Rostampour, R. (2024). Academic Challenges under the Lens of Item Response Theory: An Item Level Analysis of Measurement Precision in Perceived Academic Challenges Scale. Manuscript in preparation.

Manuscript 3: Rostampour, R. (2024). Exploring the Multidimensionality of the Perceived Academic Challenges Scale: A Bifactor ESEM Approach. Manuscript in preparation.

Acknowledgements

First and foremost, I would like to acknowledge with deep respect the Lekwungen peoples, upon whose ancestral lands the University of Victoria stands. I extend this acknowledgement to the Songhees, Esquimalt, and WSÁNEĆ peoples, whose enduring connection to the land continues to shape and enrich our community. I have been fortunate to be welcomed as a guest and learner on this land, benefiting from the rich cultural heritage and wisdom of these Indigenous communities.

This research journey would not have been possible without the financial support I received. I am profoundly grateful to the Social Sciences and Humanities Research Council (SSHRC) of Canada for their Insight Research Grant (PAR-IT; PI Allyson Hadwin, 435-2018-0440) and the multisite partnership grant (the UniVenture consortium; PI Sherry Stewart, 895-2019-1021). Additionally, I wish to express my heartfelt appreciation to the generous donors, Muriel Beverley Vaio Law Scholarships and Ian Cameron Scholarship in Educational Measurement & Evaluation, whose awards have significantly contributed to the realization of this work.

I owe an immense debt of gratitude to my supervisor, Dr. Allyson Hadwin. Her unwavering support, openness to my research interests, and guidance have been instrumental in shaping not only this dissertation but also my growth as a researcher. Her mentorship has been a beacon throughout this academic journey, inspiring me to push boundaries and explore new horizons in our field.

My sincere thanks go to my advisory board members. Dr. Phil Winne's deep and insightful comments, coupled with his careful reading of my dissertation, have not only enhanced the quality of this work but have also laid a foundation for my future academic endeavors. I am equally indebted to Dr. Stuart McDonald, whose enlightening classes on *structural equation modeling* and *multilevel modeling* ignited my passion for advanced methodological approaches. His teachings have profoundly influenced my perspective on how studying learning and self-regulation can be advanced by deploying sophisticated methodologies.

Last, but certainly not least, I want to express my deepest gratitude to my partner, Aati. Her endless support, patience, and encouragement have been the bedrock of this long and challenging process. Without her endless support, this research would not have reached its current form. Her love and support have been my constant source of strength and motivation throughout this journey.

To all those mentioned and the many others who have contributed in various ways to this work and my academic growth, I offer my sincerest thanks. This dissertation stands as a testament to the collective support, wisdom, and encouragement I have received throughout this transformative journey.

Dedication

Learning is not just about acquiring mere skills, but about the rich tapestry of experiences and interactions that shape our understanding. In that spirit, I dedicate this dissertation:

To all the teachers who have guided me on this journey of discovery and growth. Listing them here is not feasible, so I mention my first and most formative teachers, and I am sure they represent the rest of the equally loved ones: my father, who dedicated half a century to the noble profession of teaching, and my mother, a primary school teacher for three decades.

To my peers and friends, the most notable of whom is my partner Aati, whose collaboration, insights, and friendship have enriched my academic journey in countless ways.

And to all the students I have had the privilege to work with. Your curiosity, perspectives, and enthusiasm have taught me as much as I hope to have helped you learn.

Each of you has contributed to my understanding that learning is, at its core, a profoundly social phenomenon. This work is a reflection of the collective knowledge and experiences we have shared.

Chapter 1: Theoretical Framework

Education is what survives when what has been learned has been forgotten. - B. F. Skinner.

This chapter provides an exploration of academic challenges in the context of self-regulated learning (SRL). It begins by outlining the concept of SRL, focusing on the Winne and Hadwin model and its application to student success. Next, the benefits of self-report data for understanding SRL and its relation to academic challenges are discussed. The chapter then defines academic challenges from conceptual and operational perspectives, highlighting how these challenges are perceived and measured. Next, the temporal and contextual nature of academic challenges is examined, addressing how these challenges vary within individuals over time and across different situations. The Academic Challenges Scale is introduced followed by explaining the categorization of challenges into key domains of regulation, and the importance of accurately measuring these challenges is emphasized. Finally, the chapter concludes by identifying five main gaps in current measurement models, discussing how addressing these gaps can improve the measurement of academic challenges.

Self-regulated Learning (SRL)

Self-regulated Learning (SRL) refers to the self-directed process in which individuals define learning tasks, set learning goals, monitor their progression, and modify their approach to accomplish those goals (Boekaerts, 1997; Efklides, 2011; Pintrich, 2004; Schunk & Greene, 2018; Winne, 2005; Winne & Hadwin, 1998; Zimmerman, 1989, 2002). Despite theoretical differences among several SRL models, they all view students as agents capable of managing their learning through metacognitive guidance, active participation, and strategic action such as planning, monitoring, and evaluating (Winne & Perry, 2000; Wolters, 2003; Zimmerman, 1990).

Such a deliberate process involves regulation of diverse range of overt and covert behaviours toward meeting learning goals (Dunlosky & Ariel, 2011; Hadwin et al., 2011; Schunk

& Greene, 2018; Winne, 1997; Zimmerman, 2013). Consequently, SRL as a research domain is an umbrella framework under which the influence of various aspects of regulation on learning- including cognitive, metacognitive, behavioral, motivational, volitional, and affective processes - are studied and promoted (see Panadero, 2017 for a review of six models of SRL).

Another consistency across SRL models is the portrayal of goal-directed behaviour as a dynamic process comprising a few distinct yet interconnected recursive phases through which learners actively manage their learning experiences (see Panadero, 2017). These phases are further explained by describing Winne and Hadwin's model of SRL (1998), which serves as the theoretical foundation for this dissertation.

Winne and Hadwin's Model of SRL

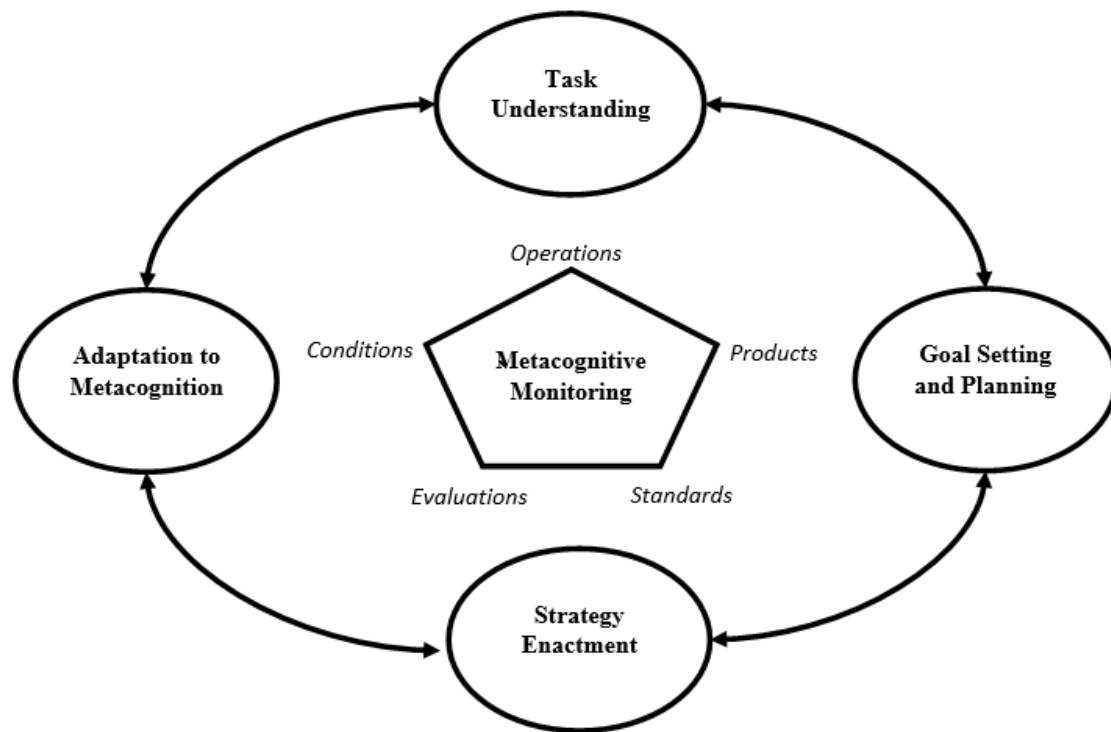
Winne and Hadwin's (1998) model of SRL explains SRL using four basic phases called (a) task understanding: generating an understanding of the task at hand; (b) goal setting and planning: forming explicit goals and making executable plans to achieve them; (c) strategy enactment: applying strategic actions; and (d) adaptations to metacognition: long-term changes to beliefs, motivations and strategies to be used in relevant future situations. This model conceptualizes SRL as a recursive process involving these phases linked in a loosely sequenced architecture. While the overall dynamics of SRL can be relatively sequential, the output of one phase can, simultaneously, feed forward, or back, to the same and/or other phases. Additionally, the prerequisite for such a multi-phase strategic control process is metacognitive monitoring which is at the core of the model (see Figure 1).

Also, Winne and Hadwin (1998) deploy another taxonomy called COPES to explain the finer-grained processes occurring within each phase. COPES stands for Conditions, Operations, Products, Evaluations, and Standards. *Conditions* are resources (affordances) and constraints

(limits) affecting learning, including both internal and external factors. Prior knowledge, beliefs, dispositions, emotional and motivational states are some examples of internal conditions while task characteristics such as time available to complete the task or task difficulty are examples of external conditions. *Operations* are cognitive processes such as Searching, Monitoring, Assembling, Rehearsing and Translating done to manipulate representations and synthesize information resulting in a product at each phase of SRL (Winne, 2013). *Products* are the information produced by operations, *Standards* are criteria against which products are evaluated, and *Evaluations* are feedback about the discrepancy between *products* and *standards* (Green & Azevedo, 2007; Winne & Hadwin, 1998). To summarize, a product is a function of an operation, evaluated against a standard and moderated by a set of internal and external conditions. The phases are linked via complex feedback loops that create a dynamics in which conditions can influence both standards, and operations and products of a phase can set conditions on other phases of SRL. In a theoretical paper discussing how the same model applies to the regulation of motivation, Winne and Hadwin (2008) emphasized that the same model can be applied to the regulation of motivational states. Figure 1 is a representation of the four phases of SRL and the COPES taxonomy.

Figure 1

The Winne and Hadwin's model of Self-Regulated Learning (1998)



Note. The four phases of SRL and the COPES taxonomy are represented. COPES taxonomy (Condition, Operation, Product, Evaluation, & Standard) at the center of the model reflects the micro processes of monitoring and control that are involved in all phases of self-regulated learning.

In line with other models of SRL, this model views regulation as a goal driven activity which is a function of two major metacognitive processes (a) monitoring: identifying discrepancies between current and desired states and (b) control: overt or covert actions to reduce the discrepancy.

Winne (1997, 2018, 2022) defines SRL as experimenting with one's own learning, involving an iterative process of IF-THEN-ELSE. The “Ifs” represent the conditions or situations that trigger self-regulated learning. This may include recognizing a learning goal, identifying a task requirement, or acknowledging a problem that needs to be solved. The “THEN” involves the actions or strategies the learner decides to implement in response to the conditions identified in

the “IF” phase. These actions are aimed at achieving the learning goal or tackling the identified challenge. This stage includes planning, selecting strategies, and applying those strategies to the task at hand. The “ELSE” component comes into play when the initial strategies and actions (the *THEN*) do not yield the desired outcome, leading to evaluation of the effectiveness of the implemented strategies and an opportunity to make adjustments by a learner. *Else* might involve trying alternative strategies, seeking additional resources, or re-evaluating the learning goal or standard. Throughout the If-Then-Else cycle, learners continuously monitor their progress and outcomes. They compare their current state with their goals and use this feedback to decide whether to continue with the current strategies or switch to new ones. The model is iterative, meaning learners might go through multiple If-Then-Else cycles within a single learning task. Each cycle helps refine their approach, leading to more effective learning strategies and better outcomes. The model highlights the adaptive nature of self-regulated learning, where learners are not rigidly following a single path but are instead dynamically adjusting their actions based on ongoing evaluations of their progress and the effectiveness of their strategies (Winne & Hadwin, 1998). In summary, the If-Then-Else model of self-regulated learning emphasizes both the conditional and adaptive nature of SRL.

To self-regulate effectively, a self-regulated learner does not merely focus on the specific task at hand. SRL prompts learners to gather knowledge about their own learning, and to experiment with how they can use that knowledge to enhance future learning. Such a focus on one’s own learning is similar to what is called meta-learning or learning how to learn in natural and artificial intelligence literatures (e.g., Wang, 2021).

Student Success from the Lens of SRL

According to Zimmerman (2015), self-regulated learning is neither an aptitude nor a mere performance skill. Instead, it is a transformative process that turns personal capacities into performance skills. Importantly, self-regulated learning is not a fixed trait but a process that can be nurtured, taught, and developed. With appropriate guidance and practice, learners can enhance the quality of their self-regulated learning (Broadbent & Poon, 2015; Cambridge, 2013; Cleary & Zimmerman, 2004; Jansen et al., 2019; Pino-Pasternak, et al., 2014; Whittlesey 2021; Zimmerman et al., 2011). Consequently, SRL goes beyond regulating study behaviours such as effective management of time and information. It is more broadly involves regulation of self by the self (see Martin 2004a, 2004b, 2007; Martin & McLellan, 2008, 2013). In other words, self is both the subject and origin of regulation (determining self), and target and object of regulation (determined Self). As goals and standards are set by the learner, student success within the context of SRL is about boosting students' autonomy and competency toward meeting self-set learning goals.

Self-regulated learning can have considerable significance for learners, primarily because it can equip them with the capabilities required to amplify their learning potential and efficiently overcome obstacles that may arise during their academic journey. By aiding learners to counter various deterministic forces that may impede their learning progression, SRL plays an essential role in promoting an agentic mode of learning in which learners actively take charge of their own learning. In this respect, SRL can be viewed as a step towards making education more inclusive, as it expands the potential of the *learning self* and assists learners in reaching their self-set learning goals.

The processes of gaining control over one's learning can increase a sense of self-efficacy (Bandura et al., 1999) and boost motivation (Pintrich, 2004). This boost of positive emotions can notably reinforce engagement with and control over learning. According to social constructivism, having a high sense of autonomy will boost students' behavioural investment in their learning, and consequently increase the likelihood of achieving the desired outcome (Reeve, 2002, 2009, 2011; Reeve et al., 2002; Su & Reeve, 2011). Sense of autonomy is associated with having a growth mindset, or holding beliefs such as: change is possible, alternatives should be considered, and effort can bring about change (Dweck, 2015, 2016; Yeager & Dweck, 2020).

Beyond encouraging agentic engagement with one's own learning, SRL also encourages a kind of learning environment that provides room for exercising agency. As an example, the perceived sense of autonomy can be supported by providing alternative choices and letting learners choose the one they prefer the most (McCombs & Marzano, 2010; Paris & Paris, 2003; Perry, 1998; Sierens et al., 2009). When task structure permits, autonomy supportive instruction has been generally found to be more effective than deterministic modes of instruction (Reeve & Cheon, 2021; Reeve & Jang, 2006; Reeve & Tseng, 2011).

Although the advantages of SRL extend beyond normative measures of academic success and align more closely with personal objectives and self-set goals, studies have demonstrated a strong association between the adoption of SRL and improved academic performance (Broadbent & Poon, 2015; Jansen et al., 2019; Pintrich, 1990; Theobald, 2021; Zheng, 2016).

Acknowledging the importance of both self-determination and adaptation to external demands in SRL leads to the conclusion that educators should not evaluate SRL success based on criteria that exclusively address one of these aspects. As an example, measures of academic performance such as grades or GPA reflect how a student performs in terms of the norms of their

institutions but do not necessarily reflect personal success. In fact, capturing a self-determined definition of success is not feasible without considering the subjective world of learners. An example of organismic-level constructs that can reflect a balance between self-determination and adaptability of one's behaviour is academic wellbeing which can be conceived as a representation of one's perceived quality of academic life (see Sfeatcu et al., 2014 for a definition of general wellbeing). Wellbeing can be viewed as a reflection of a functional balance among experiential, ought, and ideal selves. This is in line with Keyes and Annas (2009) conceptualization of wellbeing comprising of hedonic (emotional) and eudaimonic (social and psychological) dimensions. That is, feeling well and functioning well both contributes to one's sense of wellbeing. Without a balanced satisfaction of needs, and a sense of wellbeing, motivation suffers, and positive change is hardly sustainable (Deci & Ryan, 1985, 2002, 2012; Ryan & Deci, 2000; Rusk, 2022; Vansteenkiste & Ryan, 2013). Such a relationship has been found to remain valid beyond cultural, socio-economic, and personality differences (Chen et al., 2015; Church et al., 2013). Research on passion and burnout, work-life balance, and eudaimonic well-being hold similar views (Kalliath & Brough, 2008; Milyavskaya et al., 2009; Ryan & Deci, 2000; Ryff, 2013; Saville et al., 2018; Vallerand et al., 2003). Emphasizing such an organismic view of balance is an important metacognitive knowledge that learners should have and an important motive to create contexts and policies fostering a sustainable human growth through balancing self-determination and adaptation.

Similarly, research in the student success literature consistently highlights the importance of social and emotional factors for academic learning and performance (Tinto, 1982; Wong et al., 2024). The literature consistently documents the importance of social adjustment, social integration (Veldmen et al., 2019), social belongingness (Baumeister et al., 2017;

Karabenick & Berger, 2013; Karabenick & Gonida, 2018; Won et al., 2021), and social well-being (Keyes, 2005; Knoesen & Naudé, 2017).

In an era defined by the rapid turnover of knowledge and skills, the ability to learn independently has never been more crucial. Learners are increasingly faced with the need to regularly update and reinvent their skill sets, even later in life, which can be psychologically demanding. Moreover, learning is no longer confined to the walls of traditional educational settings. It has expanded beyond those boundaries, facilitated by the accessibility of online resources and platforms. From self-paced courses to YouTube tutorials, learners have a plethora of resources at their fingertips. The core processes of SRL, such as metacognitive monitoring and control, have broad applications extending beyond academia into various areas of life, offering a dynamic, personalized methodology for learning how to learn. While we cannot assure that the specific content being taught today will hold its relevance a decade from now, we can emphasize teaching a methodology that nurtures a personalized, lifelong learning. As mentioned, SRL is a comprehensive approach that spans a broad spectrum of regulatory areas, including cognitive and metacognitive strategies as well as emotional and motivational regulations. A future oriented framework of SRL, must help learners to optimally leverage future affordances, proactively navigate the ever-changing landscape of knowledge and skills, adapt effectively to changing demands, manage potential existential threats, and thrive amidst uncertainty. Such a view of self-regulation, encompassing self-determination as well as reactive/proactive adaptability, demands taking learners' subjective world into account.

The Benefit of Self-Report Data for Understanding SRL

Methods of measuring self-regulated learning can be divided into two categories. The first category involves exploring students' perceptions and recollections of their learning

behaviors. This can be done in real-time, such as with think-aloud protocols (Greene et al., 2011), or retrospectively via self-report questionnaires (e.g., MSLQ, Pintrich et al., 1993; LASSI, Weinstein et al., 2002), interviews (Zimmerman & Martinez-Pons, 1986, 1988), and closer to real-time using experience sampling methods such as study diaries (e.g., Schmitz et al., 2011). The second category involves observational methods, such as tracking learners' actions in real time or through traces of those actions during a study episode using digital traces (e.g., Winne & Baker, 2013; Winne & Hadwin, 2013; Winne, 2017) or using multimodal data (e.g., Molenaar et al., 2023; Azevedo & Gašević, 2019; Järvelä et al., 2021).

The major underlying assumption of self-report measures of SRL is that learners who engage in self-regulatory behaviours are reliably aware of what they do and can accurately recall, aggregate, and report them. While not exclusive to behavioral observation, a more pronounced underlying assumption in this approach is that certain observable patterns of self-regulatory behaviors can be used to infer the intentional and purposive implementation of those behaviors or to interpret the meaning of physiological reactions (e.g., emotions) associated with those behavioural patterns.

Self-regulated learners are believed to be metacognitive aware of their study behaviours (Winne & Azevedo, 2022; Winne, 2017). Metacognitive awareness, although not the sufficient factor for successful metacognitive control, seems to be an essential foundation upon which intentional control processes rest. In other words, metacognitive awareness is like a required steppingstone that allows learners to transform learning from a passive intake of information to an active, strategic, and ideally more efficient process (see Perry & Winne, 2006 for an indication of the plausibility of automatic regulatory processes).

Self-report measures provide a means for assessing learners' perceptions of how they feel and manage their study behavior. These perceptions, however, may not always align perfectly with their actual behaviors (Winne & Jamieson-Noel, 2002, Winne, 2004, Zimmerman, 2008). This potential discrepancy can arise due to various factors, including biases inherent to self-report measures, such as recall bias (Hassan, 2005) and social desirability bias (van de Mortel & Thea, 2008). Recall bias, which occurs when respondents do not accurately remember past events or experiences, can manifest in various ways. For instance, recency bias may occur when more recent events disproportionately influence a respondent's recall and saliency bias can arise when a single, prominent event overshadows more typical behavior in a respondent's memory. Similarly, social desirability bias occurs when respondents answer questions in a way they perceive as socially acceptable, rather than providing accurate responses.

Moreover, the accuracy of responses can also be compromised if respondents lack motivation to answer the questions. There can also be variability in respondents' abilities to introspectively access their inner worlds. Additionally, respondents may misunderstand or misinterpret the questions, leading to inaccurate responses. This can happen due to a variety of reasons - for instance, the use of items assessing more than one concept, like double-barreled items, which force respondents to evaluate two separate topics and provide a single answer. Also, items or response scales might not adequately represent the full range of the aspect being measured within a given sample. Lengthy questionnaires can also lead to response fatigue, causing respondents to rush through or not devote their full attention to later questions. While there are preventive strategies to address these potential drawbacks, it is crucial to keep these vulnerabilities in mind when designing, implementing, and analyzing self-report measures.

However, despite these potential pitfalls, self-report measures remain a valuable tool for gaining access to students' perceptions, feelings, beliefs, mental models of how they typically approach learning tasks, or how they believe learning tasks should be approached. In addition, no matter how much they deviate from objective measures, one's strategy deployment and decision making are based on their perceptions. One's perceptions can be viewed as a causal element influencing behavior construction (Bandura, 1986, 2001) and a major part of self-regulation is managing one's subjective experience whose experiential quality is only accessible to the self. Though self-representations, feelings, beliefs, needs, subjective experiences, among other personal characteristics can be measured via various methods, only self has direct access to them and more importantly experiences them.

Cross-validation among behavioral observations and self-reports should not necessarily be interpreted as evidence of the superiority of one method over the other, as they provide different information about the learning processes. Instead, rather than viewing behavioral observations and self-reports as competing methods, they can be seen as complementary. By understanding any differences between what has been observed by an observer and what has been perceived and reported by a learner, researchers can gain valuable information about how self-regulation occurs.

Transitioning from a nomothetic science of learning to a personalized and context-sensitive (e.g., idiosyncratic, idiographic) science of learning seems challenging. Difficulties may arise in applying generic knowledge about learning and studying to individual learners' needs and challenges because the interplay between learners' personal characteristics, the nature of the task, and the context creates a unique complexity. As everyone's learning strengths, weaknesses, challenges, and needs are unique, such a complexity demands considering numerous

moderators and mediators to avoid decreasing the validity and reliability of interpretations derived from nomothetic models. Also, only the individuals themselves have experiential access to their personal representations. Consequently, considering the subjective world of a learner, along with incorporating other data modalities and data collection methods, may help achieve a more personalized and nuanced understanding of how self-regulation is accomplished.

Taken together, self-report measures can provide valuable insights into perceptions, mental models, intentions, beliefs, and subjective experiences that may not be directly observable to an external observer. They are generally easy to administer, typically requiring participants to simply read and respond to statements. They are considered more cost-effective and time-efficient compared to other methods such as think-aloud protocols or interviews, especially when using online survey platforms. Self-report measures can readily be analyzed statistically, thanks to the quantitative nature of the responses. The psychometric properties of these measures can be examined in various contexts and across different samples. This standardization procedure enables consistent operationalization of constructs, facilitating the cross-validation and integration of findings from multiple studies.

Furthermore, large-scale post-secondary student success screening at institutional levels often relies on performance measures, overlooking the mediatory processes that lead to success and high performance. They typically underemphasize the importance of SRL, despite its crucial role in academic achievement. Incorporating comprehensive SRL assessments into institutional surveys has the potential for institutions to develop more nuanced student profiles, enabling them to design and implement targeted interventions that address students' specific self-regulatory needs and challenges. Such data could inform the implementation of policies and practices that foster an educational environment supporting the development of self-regulatory skills.

In addition to their measurement functions, self-report measures can also serve as self-assessment tools that can enhance metacognitive awareness among students (Panadero et al., 2016, 2017). The content of the items and underlying concepts associated with them can bring important facets of SRL into the respondents' conscious awareness. For instance, an item asking students to rate the degree to which they struggle with breaking a big learning task down into manageable steps highlights the importance of this strategy for them, reinforcing the concept in their minds.

While raising initial awareness is crucial, the larger psychoeducational goal is to help respondents transform this awareness into metacognitive knowledge. Providing survey reports to students can facilitate such a transformation. Such feedback allows respondents to accurately interpret their survey results. As an example, it can help them interpret a self-diagnostic profile that reflects their learning strengths and challenges and show them further possible steps they can take to improve their learning strategies and tackle challenges.

To conclude, self-report measures coupled with feedback can work as an effective tool for fostering metacognitive awareness and self-evaluation among students. The expectation of receiving feedback may also improve the quality of responses and mitigate biases, such as social desirability bias. When students know they will receive an individualized report based on their responses, they may be more motivated to answer accurately and honestly. However, this hypothesis requires empirical evidence to confirm its accuracy.

Academic Challenges

Despite extensive research on the factors that contribute to student success, there is limited investigation into the specific challenges students face while studying. Current findings highlight that time management, cognitive strategies usage, concentration, and boredom are the

most reported challenges among learners (Koivuniemi et al., 2017). Koivuniemi et al., (2017) found that students with higher levels of SRL skills, as indicated by the MSLQ, reported more frequent cognitive challenges during individual study sessions compared to those with lower SRL skills.

Similarly, Hadwin et al. (2019) found that the most frequently reported weekly challenges over an academic semester included motivation and emotion, goal setting and planning, and cognitive challenges. Motivation and emotional challenges were consistently high, irrespective of students self-reported rates of goal attainment. In line with Koivuniemi et al. (2017), cognitive challenges were more commonly reported by students who were academically successful as indicated by their higher rates of weekly goal attainments.

Defining Academic Challenges as a Concept, Construct, and Operationalized Construct

To define academic challenges, it is necessary to approach it at two levels: concept and construct. Concepts are generalizable properties or characteristics associated with objects, events, or people. They represent broad ideas that can vary in levels of abstraction, ranging from simple and concrete (e.g., weight) to more abstract and complex (e.g., personality). Constructs, on the other hand, are the concepts that are operationalized by theoretical models. In other words, constructs are specifically chosen or created to explain a given phenomenon within a theoretical framework. Constructs are often composed of multiple related concepts (Bhattacharjee, 2012).

Conceptual Definition

Learners engage in SRL to bridge a gap between a current state and a desired learning outcome (Dunlosky & Ariel, 2011). Challenges represent the gap where students perceive their current abilities or strategies as insufficient or requiring significant effort, making it difficult or demanding to meet their goals.

In the context of self-regulated learning, the concept of academic challenges refers to perceived discrepancies between a student's current state and desired academic goals, necessitating self-regulated learning to bridge the gap. Challenges can manifest across various domains, impacting a student's motivation, engagement, goal-setting, time management, cognitive processing, metacognitive strategies, and social-emotional well-being.

Challenges can be perceived as either demands or difficulties that can hinder progress or as opportunities for learning and growth. How a student perceives these challenges may determine whether they trigger learning and self-regulation or impede them. Factors such as the intensity and duration of the challenge, along with one's self-efficacy in tackling challenges, and their self-regulatory skills and competencies, are also critical in changing this dynamic perception.

Construct Definition

As a construct, challenge experiences are translated and organized by the theoretical model into a multidimensional structure encompassing broad regulatory challenges. In other words, perceived academic challenges are mapped onto key areas of regulation, including behavioral, motivational, volitional, cognitive, metacognitive, social, and emotional regulation.

To elaborate, the construct of academic challenges allows for integration of concrete triggers or barriers of learning (i.e., *IFs* such as context, conditions, states, etc.) into a more abstract categorization of challenges in different regulatory domains. This can facilitate strategic adjustment by helping learners and researchers to systematically view various instances that trigger or disrupt learning as indicators of broader regulatory issues.

In other words, challenges can be viewed as “what” needs to be regulated, and SRL and strategic actions as different layers of “how” regulation should be made. Understanding the

“what” (academic challenges) matters because it can provide meaning and context to the “how”.

This understanding can lead to a taxonomy where strategies to address challenges in different regulatory domains are organized. Table 1 provides definitions and examples for the key notions discussed so far.

Table 1

Definition of the Concepts and Constructs

Notions	Definition	Example
Academic Challenges (Concept)	Academic challenges refer to perceived discrepancies between a student's current state and desired academic goals, triggering self-regulated learning to bridge the gap.	Experiencing an inability to finish tasks on time, leading to incomplete assignments and poor exam preparation.
Academic Challenges (Construct)	Metacognitive appraisals of challenges (by learner or researcher) as indicators of broad regulatory domains (e.g., behavioral, motivational, volitional, cognitive, metacognitive, emotional, social)	Recognizing that one's inability to finish tasks on time falls under metacognitive challenges (e.g., failing to break tasks down, not knowing how to approach, low task understanding), which need targeted strategies to improve.
Self-regulated learning (SRL)	Self-regulated learning refers to individual learners taking metacognitive control of cognitive, behavioral, motivational and emotional conditions/states through iterative processes of planning, monitoring, evaluation, and change (Hadwin et al., 2017).	Developing a detailed study plan, including regular review sessions, self-testing, and adjusting methods based on what works best to monitor understanding of the material.
Learning strategies/techniques	the specific "methods" or "what to do" to address the challenges such as time management techniques, note-taking methods, or mnemonic devices.	Using concept mapping to visually organize and connect key concepts from the course material, making it easier to understand and remember the information.

Temporal and Contextual Nature of Academic Challenges

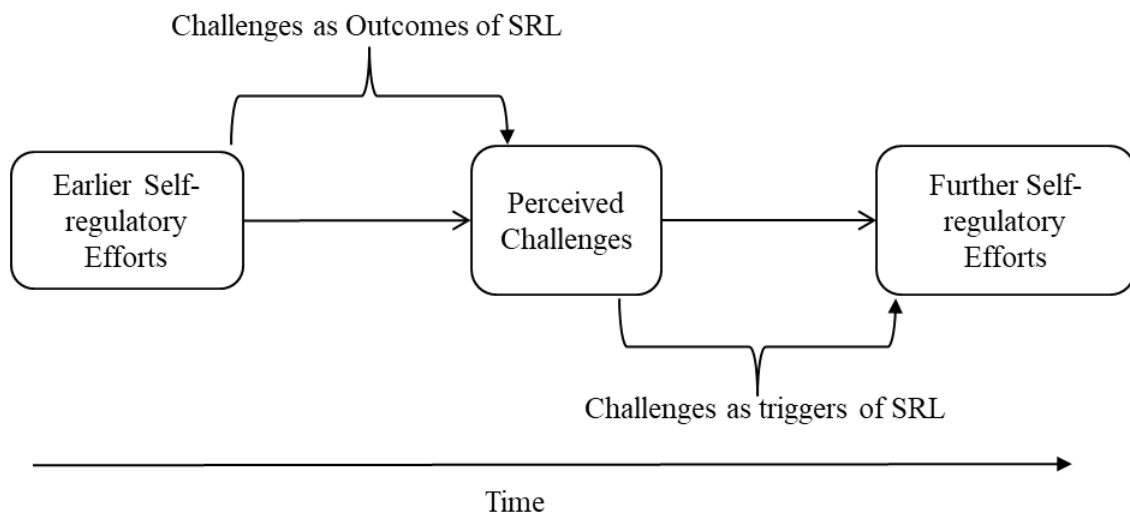
Within Person Dynamics of Academic Challenges

Using a cross-sectional design to measure different sub-processes of SRL should not be interpreted as a denial of the inherent dynamism of SRL. The processes of SRL are temporally recursive and constantly evolving. However, for the sake of practicality and efficiency in research, we often resort to creating a cross-sectional snapshot of these dynamic processes. This approach, while reductionistic, can still provide valuable insights into one's SRL. In the context of academic challenges, a cross-sectional design allows us to capture a broad overview of one's academic challenges over a particular time-window.

However, this snapshot falls short in assessing the ongoing, fluid nature of the construct of academic challenges shown in Figure 2. To clarify, academic challenges result from previous self-regulatory efforts and triggers future self-regulation cycles (Hadwin et al., 2019.; Järvelä et al., in preparation). Therefore, a cross-sectional model of academic challenges should be interpreted with the understanding that it provides a simplified representation of a complex, dynamic process, without explicitly modeling the focal dimension of any dynamic system, i.e., time.

Figure 2

A Conceptual Model Showing the Temporal dynamic of Academic Challenges



As long as the limitations are acknowledged and the quality of measurement is ensured, cross-sectional designs can still offer valuable contributions to our understanding of how one manages their learning. For instance, they provide a snapshot of a population at a particular point in time in terms of the prevalence and patterns of academic challenges. Although they cannot establish causality, cross-sectional studies are also useful for exploring associations between such a snapshot of academic challenges and other co-administered individual and contextual differences. These associations may warrant further investigation through other research designs.

Situation Specificity of Academic Challenges

SRL can be viewed both as an event, representing specific instances of self-regulatory behavior, and as an aptitude, reflecting a student's overall tendency for self-regulation (Winne & Perry, 2000). Measuring academic challenges is relevant to both perspectives, yet its utility as a self-report measure requiring students to aggregate memories across learning episodes more lies in identifying relatively consistent challenges. This view can be interpreted as accurate if three main assumptions hold: (a) stable individual characteristics: students possess relatively stable traits and (e.g., biobehavioral dispositions, Trofimova, 2021; Trofimova & Netter, 2021) and/or study habits which create some degrees of stability in self-regulatory practices and perceptions (e.g., Svartdal et al., 2022), (b) contextual stability: the academic context (e.g., disciplinary demands, school climate, etc.) remains relatively stable during the time-window of measurement, providing a kind of fairly consistent backdrop for self-regulation, and (c) limited assessment time-window: students report their challenges within a relatively short time frame (e.g., the past 2 weeks) in which individual and contextual characteristics are likely to remain stable.

However, a critical issue related to self-report measures of SRL is the lack of situational specificity in some of these measures (Winne et al., 2004). Conversely, literature supports the

significant role of context in influencing these self-reports (Hadwin et al., 2001). Individual, contextual, and temporal characteristics can alter students' perceptions of academic challenges as well as their self-regulated learning behaviours. The extent to which these characteristics are stable, and how they interact is a matter that needs empirical exploration.

Academic Challenges Scale

The Academic Challenges Scale is one of four measures included in the Self-Regulated Learning Profiles and Self-Diagnostic Tool. The other three psychometrically independent scales measure students' perceptions of the importance of self-regulatory practices (SRL-Importance), their confidence in employing these practices (SRL-Self-Efficacy), and their self-reported frequency of using self-regulatory practices (SRL-Practices). Together, these four measures offer a comprehensive and interconnected profile of students' self-regulated learning behaviors.

Presented in Appendix A, the Academic Challenges Scale was developed iteratively over a several years through data collection and pilot testing with post-secondary students. Initially, Hadwin et al. (2019) asked students to document the challenges they faced during their weekly study sessions over a semester, aiming to capture firsthand insights into their experiences. These responses were then qualitatively analyzed and compiled into a dropdown list, providing a structured method for subsequent students to identify their encountered challenges. An “*other*” option was included to allow students to contribute any unlisted challenges, ensuring the list remained comprehensive and reflective of actual student experiences. This list was continually refined and expanded until it reached saturation, indicated by a negligible number of new, uncategorized challenges reported over two consecutive semesters.

The challenges were hypothesized to fall into five principal areas of regulation crucial to student success: motivation, behavior, cognition, metacognition, and social-emotional domains

This categorization aimed to comprehensively represent the diverse academic challenges students reported. As detailed in Study 1, this initial five-factor solution was later revised to a six-factor model to better capture the complexities of the reported challenges. Further details about the scale will be provided in Chapter 3.

Categorization of Academic Challenges

The multidimensional modeling of academic challenges as a taxonomy of perceived self-regulatory demands is driven by a combination of psychological and pragmatic considerations. As mentioned, the theoretical construct categorizes challenges into underlying regulatory domains that are linked to student success (Bowden et al., 2019; Wong et al., 2024). Although these regulatory dimensions are interlinked, a component-based modeling of them seems beneficial because research in these areas comprises related yet distinct disciplinary literatures. These domains include behavioral, motivational, volitional, cognitive, metacognitive, emotional, and social areas of regulation. In fact, a learner's profile of challenges serves as a map representing their struggles across the main regulatory dimensions that impact learning and self-regulation in various contexts. Such a personal profile can provide a context that helps tailor strategies and supports more precisely to students' needs.

In summary, mapping challenges onto key regulatory domains aligns with the disciplinary nature of research in these areas and provides a practical framework for organizing targeted interventions, strategies, and supports for each regulatory domain. Table 2 includes a brief overview of the academic challenges dimensions.

Table 2*Breakdown of Challenges in Each Regulatory Domain*

	Definition	Examples
Motivational Challenges	Struggles with willingness and drive to engage in academic tasks.	Lack of interest in the subject matter, difficulty seeing the relevance and value of the material, a low sense of self-efficacy.
Challenges in Initiating and Sustaining Engagement	Struggles with starting and maintaining involvement in, as well as focus on, academic activities.	Struggling with Procrastination, with maintaining focus during study sessions.
Goal and Time Management Challenges	Struggles with effective time and task management	Struggles with scheduling, establishing study habits, meeting deadline.
Cognitive Challenges	Struggles with the cognitive demands placed on students in processing, retaining, and applying academic content	Difficulty comprehending complex material, elaborating and connecting concepts.
Metacognitive Challenges	Struggles with higher-order thinking processes that involve planning, monitoring, and evaluating one's own learning.	Struggles with assessing one's own understanding, or with poor adjustment of strategies based on performance feedback.
Social and Emotional Challenges	Struggles with the ability to interact with others within the academic context and with managing emotions and maintaining wellbeing in the learning context.	Difficulties in finding enjoyment in the academic context, and challenges in forming positive relationships with peers and instructors.

Importance of Measuring Academic Challenges

Measuring academic challenges provides a profile of students' academic challenges within key areas of regulation. This profile can serve multiple purposes, described below.

Self-Assessment and Self-Reflection. As discussed earlier, a critical aspect of SRL is the development of metacognitive awareness, knowledge, and skills (Winne & Hadwin, 1998). In addition to developing an understanding of the learning processes and task requirements (task understanding), SRL requires learners to develop an awareness of their personal preferences, capabilities, strengths, and weaknesses (self-understanding) by engaging in metacognitive monitoring of learning within and across their study sessions (Winne & Perry, 2000). These metacognitive self-evaluations influence students' self-regulatory decisions (de Bruijn-Smolters et al., 2016; Schunk & DiBenedetto, 2016; Sitzmann & Ely, 2011). Such a self-understanding guides individual learners to take metacognitive control of their cognitive, behavioral, motivational, and emotional conditions/states through iterative processes of planning, monitoring, evaluation, and adaptation (Hadwin et al., 2017). Consequently, having both an accurate self-understanding and a clear understanding of the task (and the interplay between them) is important for effective self-regulated learning.

The Academic Challenges Scale can be used by students as a self-assessment and self-reflection tool, helping them better interpret their struggles as indicators of broader self-regulatory challenge categories. Coupled with feedback and psychoeducational guidance, this understanding can aid in deploying strategies and seeking support to effectively tackle the challenges. This process encourages the recognition of challenges as opportunities for fine-tuning and improving learning rather than viewing them as impediments.

Diagnostic Tool. The scale can function as a diagnostic tool, enabling educators to identify struggling students and customize interventions and support systems to improve their

self-regulatory capacities and learning. If challenges are not addressed strategically and timely, they can become prolonged and intense, significantly hindering a student's ability to learn effectively. This can lead to a cycle of academic struggles and decreased performance and poor wellbeing. Hence, the scale can be used for early identification of struggling students to enhance retention and performance.

The scale can also be used to demonstrate the profile of challenges across a cohort of students or within a course, leading to instructional adjustments based on students' needs. Understanding students' regulatory challenges can promote responsive instructional designs. Teachers can use information about these challenges to inform their instructional practices. For instance, a course in which students report predominantly cognitive challenges may benefit from instructional adjustments that focus on enhancing comprehension. Conversely, a course where students face motivation challenges may require strategies to boost engagement and interest, such as integrating more interactive practices. This situational specificity can be further narrowed down to the task level, allowing for precise adjustments tailored to individual learning tasks.

Research Utility. The scale can assist researchers in organizing learning strategies into taxonomies, systematizing which strategies can address specific challenges within the main areas of regulation. Translating it into Winne's IF-Then-Else model, discussed earlier, educators and counselors can organize interventions and strategies (Then, Else) based on the challenge categories they address (IFs). Additionally, the scale can be used as a construct mediating the effect of interventions on engagement, achievement, retention outcomes, and academic well-being. It offers opportunities to examine the nuanced impact of such interventions on remediating academic challenges associated with poor academic performance and well-being.

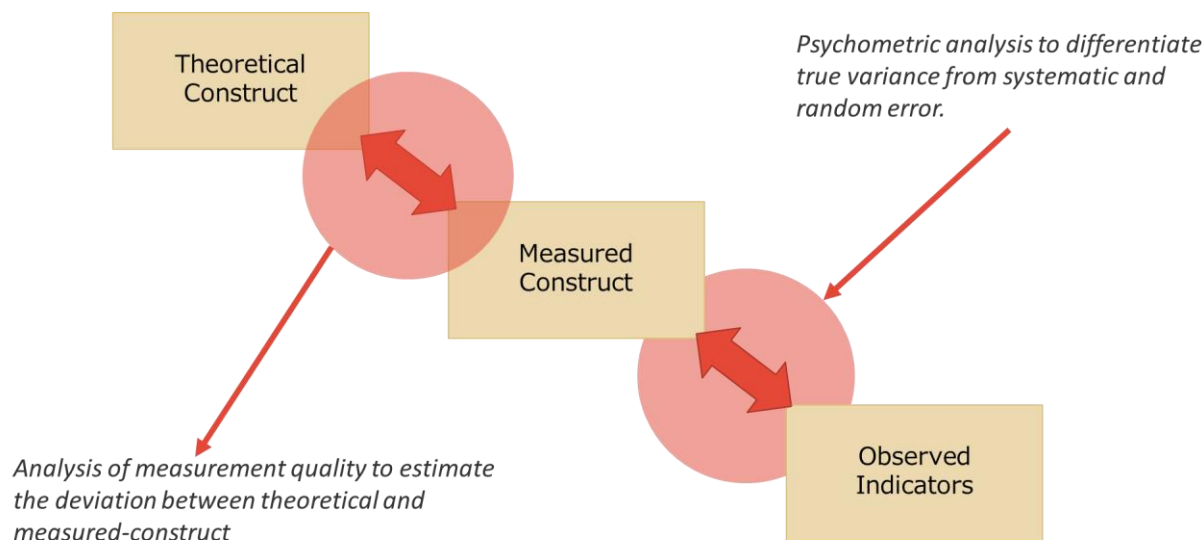
Identifying Gaps in How Measurement Models for the Academic Challenges Have been Operationalized

Psychometric validation of a self-report scale often involves investigation of the construct validity and reliability of a scale. *Construct validity* refers to the extent to which a scale accurately measures the theoretical construct it is intended to measure. This includes several subtypes: *content validity*, which ensures the scale covers all aspects of the concept; *convergent validity*, which demonstrates that the scale correlates well with other measures of the same construct; *discriminant validity*, which shows that the scale does not correlate too highly with measures of different constructs, indicating that the scale is measuring something distinct; and *criterion-related validity*, which examines whether the scale's scores are correlated with relevant outcomes or behaviors, including predictive and concurrent validity. *Reliability*, on the other hand, refers to the consistency and stability of the scale's scores over time and across different conditions. This encompasses *internal consistency*, which assesses whether the items on a scale consistently measure the same construct; *test-retest reliability*, which examines the stability of the scale over time; *inter-rater reliability*, which measures the consistency between different raters; and *parallel-forms reliability*, which evaluates the consistency of results between two equivalent forms of a test (Furr, 2021; Rust & Golombok, 2014).

The psychometric validity of a measure can further be supported by establishing measurement invariance, which refers to the degree to which a scale measures the same construct across different groups or conditions. Establishing measurement invariance ensures that comparisons made between different groups (such as gender, age, or cultural groups) are valid and meaningful. Without measurement invariance, differences in scores might reflect variations

in how different groups understand or respond to the items, rather than true differences in the construct being measured. (Milfont & Fischer, 2010; Meredith, 1993).

However, common approaches to validating scales, such as Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), often fall short in modeling certain aspects of the theoretical constructs. These methods typically rely on restrictive assumptions which may not adequately capture the complexity and nature of many psychological constructs, as will be discussed in greater detail later. As a result, the measurement models derived from these analytic methods can deviate from the theoretical constructs they are intended to represent. This misalignment occurs because the analytic methods may not fully accommodate the nuanced interactions and structures inherent in complex constructs, leading to an oversimplified representation of the construct (Marsh et al., 2005). Consequently, it is crucial to consider alternative modeling techniques to potentially improve the ecological validity of the measured constructs, ensuring that the scale accurately reflects the theoretical constructs. Figure 3 depicts the bidirectional process of translating observed indicators into measured constructs and theoretical constructs, and vice versa. The observed indicators are the actual items in the questionnaire used to measure the construct. Psychometric analysis of *item-scale* associations (i.e., bottom right arrow) is crucial to differentiate true variance from systematic and random errors. Systematic errors are consistent identifiable biases that can skew measurements, while random errors are unpredictable variations that add noise to the data. By addressing these errors, researchers can enhance the reliability and psychometric validity of their measurements, ensuring that the observed indicators accurately represent the measured constructs, and thereby, the theoretical constructs.

Figure 3*Bidirectional Process of Construct Operationalization*

The theoretical construct represents the abstract concept that researchers aim to quantify. Through the development of scales or instruments, this theoretical construct is operationalized into a measured construct. An analysis of measurement quality at the stage represented by the top right arrow in Figure 3 helps estimate any deviations between the theoretical construct and its measured proxy, ensuring the measured construct accurately reflects the theoretical idea. Methodologically, measurement error should not be equated to *item-scale* residuals but should also include the deviation from theoretical construct to measured construct.

To elaborate, measurement error is often defined as the amount of variability in observations that is unique to them and left unexplained by the commonalities among all items reflected in the latent construct. However, latent constructs are merely proxies of conceptual constructs (Hair et al., 2017). Just as residuals represent the amount of noise in constructing latent constructs from observations, there is always another types of uncertainty between conceptual constructs and their operationalized proxies. The challenge is that this latter type of error cannot easily be modeled or computationally analyzed. However, raising awareness about

ecological validity, triangulating across measurement approaches, cross-validating with multimodal data, and employing mixed-method designs that combine qualitative and quantitative analyses could provide information about the magnitude of this type of measurement error.

This bidirectional relationship in Figure 3 highlights the iterative nature of refining measurement to better capture the theoretical constructs. Improving a measure involves a dynamic interplay between (a) item revision, (b) improving measurement models, and (c) potentially modifying the theoretical construct. This process is dialectic, with the measured construct and theoretical construct mutually refining each other. Feedback from empirical data informs updates to the theoretical framework, while theoretical advancements guide the development of more accurate measurement items. This continuous cycle ensures both constructs evolve together, leading to more precise and reliable theory and measurements.

Deviation between the theoretical and measured constructs can partly be due to the limitations of the analytical methods used to operationalize measurement. Addressing these limitations can potentially improve the measurement by reducing the gap between theoretical and measured construct. In the remainder of this section, some of the limitations of common construct validation processes relevant to the measurement of academic challenges are discussed.

Need for Mapping General and Specific Components in Academic Challenges

Understanding and accurately measuring academic challenges faced by students is crucial for developing effective interventions and support mechanisms. Modeling both general and specific factors within the Academic Challenges Scale provides a more nuanced understanding of these challenges. This section discusses the importance of such an approach in modeling academic challenges.

Academic challenge areas are often interrelated, indicating high convergent validity. That is partly because experiencing a challenge in one area can often lead to challenges in other areas. For instance, a student struggling with cognitive challenges may also face emotional challenges due to frustration or anxiety. Procrastination serves as another illustrative example, as it can stem from various sources, such as motivational, cognitive, metacognitive, social, emotional, behavioral, and volitional challenges. A student might procrastinate due to a lack of interest (motivational), difficulty understanding the material (cognitive), poor planning skills (metacognitive), peer influence (social), anxiety (emotional), ineffective study habits (behavioral), or an inability to maintain focus (volitional). These interconnections suggest that academic challenges cannot be viewed in isolation, as they often overlap and influence each other.

High associations between different academic challenges can also be interpreted as an indicator of low discriminant validity, suggesting that they fall within a general challenge dimension. This means that specific challenges may not be distinct enough, and they may instead be indicators of a broader general academic challenge factor. High associations between challenges dimensions raises the question of whether academic challenges can be modeled as a unidimensional general construct rather than as a multidimensional construct comprising specific factors.

A multidimensional model measures different dimensions of a construct without necessarily summarizing them into a single overarching construct. For example, the Big Five personality model operationalizes personality into five distinct dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism (McCrae & Costa, 1997).

However, to model personality, we need additional information on how to integrate the scores of these dimensions into a comprehensive representation of the main construct of personality.

Similarly, in the context of academic challenges, a multidimensional model would identify and operationalize specific challenges such as motivational, cognitive, metacognitive, social, emotional, behavioral, and volitional factors. Each of these dimensions can be measured and analyzed separately to understand the specific areas where students face challenges. However, to fully understand the overall construct of academic challenges, we must also consider whether these specific factors can be combined to represent a general academic challenge factor.

While a unidimensional model of academic challenges simplifies the measurement and provides a practical representation of the general academic challenge, it overlooks the specific nature of challenges. Knowing the specific areas where students face challenges provides valuable information for targeted support. For example, interventions designed to improve study habits will differ significantly from those aimed at reducing anxiety or improving social engagement. Therefore, identifying and modeling specific academic challenges is beneficial for providing effective and tailored support to students. Therefore, modeling both general and specific factors is necessary due to the different questions they can address.

Common approaches to measurement often model constructs either as unidimensional or multidimensional, without integrating both general and specific factors together. Given the coexisting nature of specific and general factors, a unidimensional model of academic challenges that neglects specific challenges can produce biased estimations of these factors (Reise, et al, 2013). Similarly, a multidimensional model that overlooks the general factor can also lead to biased parameter estimates (Rodriguez, et al., 2016). Therefore, it is crucial to model both

general and specific factors simultaneously to obtain an accurate and comprehensive understanding of academic challenges. Ideally, such a model should allow for the coexistence of psychometrically non-redundant general and specific factors within a single framework. Such a model can reduce the biases of ignoring either general factor or specific factors and enable researchers and educators to respond to questions related to both the overall level and the specific nature of academic challenges in a single measurement model.

Need for Capturing Item Interdependencies in Measurement Model of Academic Challenges

Classic confirmatory factor analysis models each item in a self-report questionnaire as being exclusively caused by the psychological factor to which it belongs. In other words, any associations between an item and other existing factors in the scale are assumed to be zero. For instance, the variance of the item *"Understanding learning materials and concepts"* from the Cognitive Challenges scale is assumed to be solely explained by the cognitive latent factor as well as residuals. This restrictive assumption creates a reductionistic proxy of the theoretical construct and reduces its ecological validity. Given the interconnected nature of psychological indicators, such assumptions are overly simplistic. To exemplify, the item *"Understanding learning materials and concepts"* can have cross-loadings with items from other scales, as exemplified in Table 3.

Traditional confirmatory measurement models, which restrict cross-loadings, fail to capture these nuanced relationships. Beyond being reductionistic, suppressing these interconnections among items results in biases in the other parameters of the model (Asparouhov & Muthén, 2009; Marsh et al., 2014; Morin et al., 2020; Zhang et al., 2023). While exploratory factor analysis allows for freely estimation of these cross-loadings, it does not permit testing a

priori theoretical models. This limitation highlights the need for more advanced modeling techniques that can account for such cross-loadings while also allowing for the testing of *a priori* theoretical models.

Table 3

Potential Cross-Loadings and Explanations for an Item from Cognitive Challenges (Understanding learning materials and concepts)

Potential Cross-Loading Items from Other Scales	Explanation
Believing I can do my work (Motivation Challenges)	A student who struggles with understanding learning materials may also have low self-belief in their ability to do the work, showing an overlap between cognitive and motivational challenges.
Knowing how to tackle tasks (Metacognitive Challenges)	Difficulty in understanding materials can be linked to not knowing how to approach tasks, reflecting an overlap between cognitive and metacognitive challenges.
Putting things off (procrastinating) (Initiating & Sustaining Engagement Challenges)	Struggling with comprehension can lead to procrastination, demonstrating an overlap between cognitive and engagement challenges.
Meeting deadlines (Goal & Time Management Challenges)	Students who find it hard to understand learning materials may also struggle with meeting deadlines, indicating an overlap between cognitive and time management challenges.
Managing my emotions/feelings (Social & Emotional Challenges)	Cognitive difficulties can lead to frustration and anxiety, indicating an overlap between cognitive and emotional challenges.

In summary, the interconnected nature of academic challenges necessitates the use of modeling techniques that can accommodate the network of associations among the items and provide a more ecologically valid representation of the theoretical construct. Restricting these associations leads to a loss of important information and potential biases, underscoring the need for more sophisticated approaches to accurately capture the interconnectivities between indicators of academic challenges.

Need for Validating Arithmetic-Based Scale Scores for Practical Use

Measurement validation using advanced analytic techniques is crucial for ensuring the accuracy and appropriateness of scales for research purposes. This is facilitated by the

availability of sophisticated software and the necessary expertise to operationalize these scales effectively. However, educators often need to compute scales using simpler arithmetic methods such as summing or averaging. Evidence from factor analytic validation studies cannot be directly taken as support for the validity of interpretations of such simplified scores. Therefore, separate validation efforts are necessary to specifically validate these scores when simpler computational methods are used. This ensures that the scales remain reliable and valid for practical educational applications.

Need For an Item-Level Analysis of Academic Challenges Scale

While factor analytic approaches provide some information on individual item characteristics, the main objective and focus of these studies are at the level of latent constructs and overall model fit (e.g., factorial structure and multidimensionality). A more granular analysis to improve the scale involves performing an item-level analysis. This approach examines the performance of each individual item within the scale to identify their specific strengths and weaknesses in contributing to the overall quality of the measurement instrument.

Item-level analysis allows researchers to explore the characteristics of each item, evaluating how well each item measures the intended construct, and how it interacts with other items in the scale. By scrutinizing items individually, researchers can identify items that do not perform well, such as those that fail to differentiate between respondents with varying levels of the underlying attribute (i.e., academic challenges) or those that are potentially ambiguous or misunderstood by respondents. Items that do not effectively differentiate between respondents at different levels of the underlying construct can weaken the overall scale. For instance, an item might refer to a challenge that is too frequent or too rare, resulting in a lack of variation in responses and reducing the item's ability to reliably contribute to measuring the construct.

Items are not isolated entities; they interact with each other to form a cohesive scale. If an item does not align well with other items, it can disrupt the factorial structure of the scale. This misalignment can be due to conceptual overlaps or differences not accounted for during scale development. Additionally, items that are interpreted differently by different respondents can introduce noise into the measurement. Through item-level analysis, researchers can identify items that might function differently across diverse groups of respondents, revealing such potential biases. This is crucial for ensuring that each item is valid for any random respondent from the intended population.

Moreover, item-level analysis allows for evaluation of the response options used for the scale. Researchers might find that different items benefit from varying levels of granularity in response options. Just as an example, items related to emotional challenges might require a more nuanced response scale to capture subtle differences in feelings, whereas items related to time management might be sufficiently captured with fewer response options.

Refining a scale through item-level analysis involves making targeted revisions based on the performance of individual items. This might include rewriting items, adding new items to cover gaps, or removing poorly performing items. By complementing factor analytic methods with item-level analysis, researchers can gain a more comprehensive understanding of how each building block of the scale contributes to its overall effectiveness.

Need for Assessing Measurement Precision Across Different Levels of Academic Challenges

Classic factor analytic methods assume that the quality of measurement is constant across all levels of the attribute being measured (e.g., academic challenges). However, a scale may reliably measure average levels of a construct but fall short in measuring high and/or low levels

of that construct. Analytic methods that allow for evaluating different levels of measurement precision across the spectrum of the attribute are needed, particularly when the scale has diagnostic objectives, such as the academic challenges scale.

Factor analysis produces parameter estimates that are based on sample statistics. For producing norm-based scores, a sample invariant analytic method is needed to allow for the identification of generalizable cut-off scores. This is particularly important in diagnostic scales, where consistent measurement precision and reliable cut-off scores are essential for accurate diagnostics and comparisons across diverse groups.

Summary of the Chapter

Deviation between the theoretical and measured constructs can partly be due to the limitations of the analytical methods used to operationalize measurement. Addressing these limitations can potentially improve the measurement by reducing the gap between theoretical and measured construct. Five gaps in the measurement modeling of psychological constructs, such as the academic challenges, using common measurement validation practices were discussed. These gaps highlight the need for (a) mapping general and specific components, (b) capturing item interdependencies in the measurement model, (c) validating arithmetic-based scores for the scale's practical use, (d) conducting item-level analysis of measurement quality, and (e) assessing measurement precision across different levels of the Academic Challenges.

Needs (a) and (b) can be addressed by adopting measurement models with an analytic focus on the multidimensional structure of the scale. An ideal analytic model would be capable of handling both (a) and (b) simultaneously, providing a more aligned representation of the theoretical construct than each separately.

In addition to choosing less-reductionistic approaches, the remaining gaps (c, d, e) also require an analytic focus on the items constituting the academic challenges scale. The methodologies used to address these gaps will be explained in Chapter 2.

Chapter 2: Methodological Considerations

"You can't manage what you can't measure" - Peter Drucker

The process of measurement is a cornerstone in all psychological research and thus, warrants more rigorous examination than it commonly receives. The discrepancy between theoretical constructs and their measured counterparts can be partially attributed to limitations in the analytical methods used for measurement operationalization. Improving these methods may help reduce this gap. Key areas for improvement in measuring psychological constructs like academic challenges were identified in Chapter 1. In this chapter, the methodologies employed to address these identified needs and mitigate the specified measurement gaps will be delineated and discussed. Table 4 highlights key areas for construct measurement improvement and provides corresponding examples of analytic methods that can be employed to address each area. Additionally, it includes the main headings under which these analytic approaches are explained in this chapter.

Table 4

Key Objectives for Improving Construct Measurement in Academic Challenges Scale

Objectives	Analytic Methods	Section Headings
(0) Validating Construct Using Common Approaches	-Exploratory Factor Analysis -Confirmatory Factor Analysis	Exploring the Multidimensionality of the Academic Challenges Scale
(1) Operationalizing both global and domain-specific constituents of the construct	-Bifactor Measurement Model -Higher-order Measurement Model	
(2) Integrating cross-loadings within the measurement model	Exploratory Structural Equation Modeling (ESEM)	
(3) Operationalizing both global and specific constituents of the construct AND integrating cross-loadings within the measurement model	-Bifactor ESEM Model - Higher-order ESEM Model	
(4) Conducting granular analyses of psychometric properties at the item level	Item Response Theory (IRT)	Zooming In: Moving from Scale Multidimensional Structure to Item-Level Precision
(5) Assessing the precision and stability of measurement across the continuum of Academic Challenges	IRT (e.g., Graded Response Model)	
(6) Empirically validating composite scores derived from arithmetic aggregation for applied contexts	-Item Response Theory -Classical Test Theory (CTT)	
(7) Validating the Scale by examining how it decomposes populations into meaningful clusters	Person-Centered Analysis (Latent Profile Analysis)	From Variables to Individuals: Complementary Person-Centered Validation of the Academic Challenges Scale

While the first objective provides the foundational validation necessary to set the stage for further analysis, objectives 1-3 are aimed at improving the modeling of the multidimensional structure of the academic challenges scale. This involves operationalizing both global and specific constituents of the construct and integrating cross-loadings within the measurement model by using sophisticated analytic methods such as Bifactor and Higher-order Measurement Models within CFA and Exploratory Equation Modeling frameworks.

Objectives 4 and 5 shift the focus to a more granular level, examining individual items within the scale. This item-level analysis is essential for understanding the performance, precision, and stability of measurement across the continuum of Academic Challenges, highlighting areas where further item refinement is needed. Investigating the reliability of measurement at different levels of the academic challenges is important given the diagnostic aims of the scale. This item-level analysis complements the broader multidimensional analysis, offering a more complete view of the scale's measurement quality.

Objectives 6 and 7 serve as important extensions to the validation process. Objective 6 focuses on empirically validating composite scores derived from arithmetic aggregation for practical contexts, ensuring these scores are meaningful and reliable for applied use. Objective 7 calls for a complementary person-centered validation of the scale. Examining how the population is clustered into meaningful groups using the scale provides valuable information about how the scale can be used to identify struggling students. The explanation of these objectives and their corresponding analytic methodologies will follow the three main section headings presented in Table 4.

Exploring the Multidimensionality of the Academic Challenges Scale

Common Approaches to Construct Validation

The psychometric validation of self-report scales encompasses an examination of construct validity and reliability. *Construct validity*, which assesses the extent to which a scale accurately measures its intended theoretical construct, comprises several subtypes: *content validity*, *convergent validity*, *discriminant validity*, and *criterion-related validity*. These subtypes collectively ensure that the scale adequately covers all aspects of the theoretical construct, correlates appropriately with related measures, distinguishes itself from measures of different constructs, and demonstrates relevance to related outcomes or behaviors (DeVellis & Thorpe, 2021; Raykov & Marcoulides, 2011).

Reliability pertains to the consistency and stability of scale scores across temporal and contextual dimensions. This includes *internal consistency*, *test-retest reliability*, *inter-rater reliability*, and *parallel-forms reliability*. These reliability indices evaluate the scale's consistency in measuring a single construct, its stability over time, agreement between different raters, and consistency across equivalent forms of the test, respectively (Furr, 2021; Rust & Golombok, 2014).

A crucial aspect of scale validation is the establishment of *measurement invariance*, which evaluates the consistency of a scale's measurement properties across diverse groups or conditions. This property ensures that observed score differences reflect genuine variations in the construct rather than disparities in item interpretation or response patterns among different populations (Milfont & Fischer, 2010; Meredith, 1993). Similarly, the globalization of psychological research necessitates greater attention to cross-cultural validity and adaptation processes. This includes considerations of linguistic and cultural equivalence in translated

measures, ensuring that scales maintain their psychometric properties across diverse cultural contexts (van de Vijver & Tanzer, 2004).

Exploratory and Confirmatory Factor Analysis

Exploratory Factor Analysis (EFA) is a statistical technique commonly employed in the initial stages of scale development or when the underlying structure of a set of variables is unknown. EFA aims to uncover the latent factor structure by examining the patterns of correlations among observed variables without imposing a predetermined model. In EFA, researchers allow the data to guide the identification of factors, making it particularly useful when there is limited theoretical basis for specifying the number or nature of factors in advance. However, this data-driven approach means that an *a priori* model cannot be explicitly tested in EFA. The lack of a prespecified model can lead to challenges in interpretation and potential capitalization on chance variations in the data, necessitating cross-validation of the resultant factor structure (Costello & Osborne, 2005; Fabrigar et al., 1999).

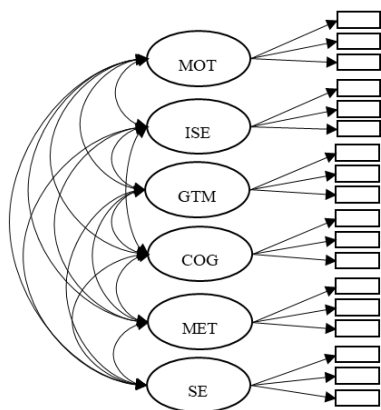
In contrast, a Confirmatory Factor Analysis (CFA) represents a theory-driven approach that allows researchers to test specific hypotheses about the underlying structure of a set of variables. This method is particularly valuable when there is a strong theoretical or empirical basis for specifying the number of factors and their relationships *a priori*. A CFA model can provide several important pieces of information. First, it offers estimates of factor loadings, indicating the strength of the relationship between each observed variable and its hypothesized latent factor. Second, it provides estimates of the correlations between factors, which can offer insights into the distinctiveness or overlap of the measured constructs. Third, a CFA yields various fit indices that allow researchers to assess how well the hypothesized model fits the observed data. These may include the Chi-square statistic, Comparative Fit Index (CFI), Tucker-

Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA), among others. Additionally, modification indices can suggest potential improvements to the model, although these should be considered cautiously and in conjunction with theoretical considerations (Brown, 2015, Kline, 2015).

The most common Confirmatory Factor Analysis (CFA) model in psychological research is the correlated-factor CFA (Figure 4). In this model, a construct is conceptualized as comprising multiple dimensions that are interrelated. Each dimension is represented as a latent factor, and these factors are allowed to correlate with one another. This correlated-factor approach is particularly useful in cases where theory suggests that a construct has distinct but related components, as it can provide insights into both the uniqueness and interrelatedness of these components (Brown, 2015; Kline, 2015).

Figure 4

Correlated Factor CFA for the Academic Challenges Scale



Note. The diagram is a path diagram commonly used in structural equation modeling (SEM). The ovals (MOT, ISE, GTM, COG, MET, SE) represent latent variables, which are the 6 theoretical dimensions of academic challenges, namely, Motivation, Initiating & Sustaining Engagement, Goal and Time Management, Cognitive, Metacognitive, and Social & Emotional challenges. The rectangles linked to each latent variable represent observed variables or measured items. To make the figure more concise, only three items per scale were represented. The lines (arrows) connecting the latent variables to the rectangles indicate the relationships between the constructs and their corresponding observed measures. The double-headed arrows between the latent variables suggest correlations or covariances between these constructs.

For example, in a correlated-factor CFA, dimensions of academic challenges scale (e.g., motivation, cognitive, metacognitive, etc.) would be modeled as separate but potentially correlated latent variables. Such a structure allows researchers to examine the relationships between different aspects of a construct, assess the discriminant validity of its dimensions, and evaluate the overall fit of the proposed multidimensional structure to the observed data.

However, conventional validation methodologies, such as EFA and correlated CFA, often prove inadequate in capturing the full complexity of psychological constructs. These methods typically rely on restrictive assumptions that may oversimplify the intricate nature of the constructs under investigation. Consequently, the resultant measurement models may deviate from the theoretical constructs they aim to represent, potentially compromising the ecological validity of the measured constructs (Marsh et al., 2005). Put differently, these analytical methods impose certain assumptions that may render them unsuitable for assessing the quality of measurement for specific real-world phenomena. In the following subsections, I examine the limitations of correlated CFA models in operationalizing the academic challenges and introduce methods that offer more sophisticated approaches to modeling the multidimensional structures of this construct. These methods potentially enhance the alignment between theoretical constructs and their operationalization in measurement instruments.

Modeling a Latent Construct of General Academic Challenge

A correlated measurement model can be conceptualized as a representation of the dimensions of a construct rather than the construct itself. To illustrate this concept, consider the widely recognized Big 5 personality model (Costa & McCrae, 1992; Goldberg, 1993). In a correlated CFA of this model, the dimensions of personality (e.g., neuroticism, extraversion) are represented, but not personality as a holistic construct. To fully operationalize personality, the

measurement model should clarify how these dimensions collectively represent the overarching construct.

Additionally, in a correlated CFA model, inter-factor correlations are estimated as components of the measurement model. However, these correlations are not operationalized as elements that can be used in the structural model for hypothesis testing purposes. For instance, in a six-factor model of academic challenges, the associations between factors are operationalized as 15 bivariate correlations. Nevertheless, this measurement model cannot inform what these covariations collectively represent, nor can it determine whether they reflect a general factor embodying the construct itself.

From a statistical perspective, it is feasible to incorporate a general factor into any model comprising several correlated first-order factors. However, computational feasibility is merely a necessary, not sufficient, condition for modeling a general factor. Conceptual and theoretical considerations are paramount in substantiating whether and how the covariation among items and/or factors can be modeled as a general factor.

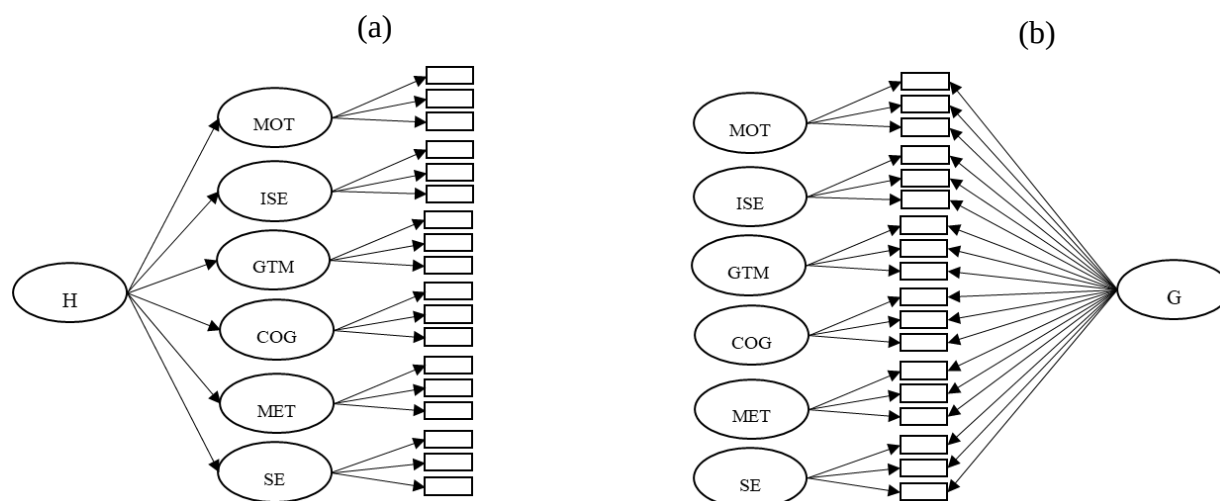
It is important to note that failing to model a potential overarching general factor may result in inflated estimates of factor correlations and regression coefficients (Asparouhov & Muthén, 2009). This inflation can lead to multicollinearity issues when attempting to use these factors as parallel predictors of potential outcomes.

Two principal approaches within the domain of common factor models, known as higher-order and bifactor models (Figure 5), are frequently employed to investigate the possibility of modeling an overarching latent construct (Brunner et al., 2012; Rindskopf & Rose, 1988). Both models incorporate a set of specific factors and typically one general factor, but they differ in

their structuring of relationships among the factors and the observed variables. The subsequent subsection will provide a detailed overview of these two modeling approaches.

Figure 5

Higher order (a) and Bifactor (b) Measurement Models for the Academic Challenges



Note. H: Higher-order Factor; G=General Factor

Higher-order Measurement Model and Bifactor Measurement Model

Higher-order Measurement Model. In a second-order factor model, the hierarchical structure of latent constructs is represented through a two-tiered approach. The observed variables (items) are modeled as direct indicators of the first-order specific factors, which in turn serve as indicators of a higher-order general factor (Figure 5a). This configuration allows for the parsing of variance at multiple levels: the specific factors account for the shared variance among their respective observed indicators, while the general factor captures the common variance among the specific factors (Rindskopf & Rose, 1988). This hierarchical structure provides a parsimonious representation of complex multidimensional constructs, allowing researchers to model both the specificity of individual dimensions and their shared variance as a higher order

factor. The higher-order factor model is particularly useful when theory suggests that a higher-order construct underlies and explains the correlations among lower-order factors.

Bifactor Measurement Model. The bifactor model (5b) offers an alternative approach to modeling multidimensional constructs. In this model, each observed variable is simultaneously influenced by both a general factor and a specific factor. In contrast to the higher-order model, the bifactor model allows the general factor to directly influence all observed variables, while simultaneously accounting for the effects of the specific factors. This configuration enables researchers to partition item variance into that attributable to the general factor and that unique to specific dimensions (Chen et al., 2006). Consequently, the bifactor model is particularly advantageous when researchers wish to examine the relative contribution of general and specific factors in explaining item variance. It also facilitates the examination of how external variables differentially relate to the general factor and specific factors, providing a more nuanced understanding of construct-criterion relationships (Reise, 2012).

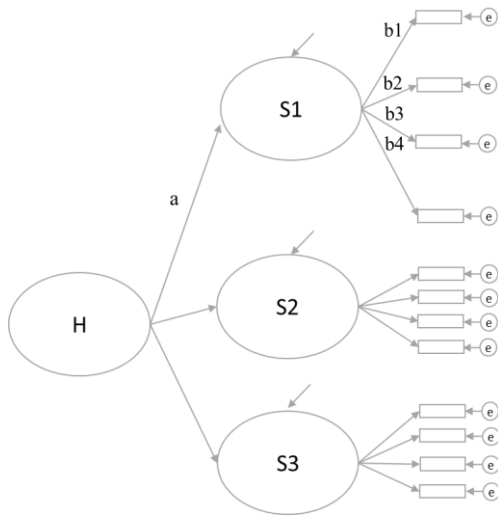
Contrasting Higher-order and Bifactor Models. Both models offer distinct advantages and limitations. The choice between them often depends on theoretical considerations, the nature of the construct being studied, and the specific research questions at hand.

In a higher-order model, the general factor's influence on observed variables is channeled through the intermediary layer of specific factors. As illustrated in Figure 5a, the higher-order factor model demonstrates an indirect relationship between a general factor and observed items. This indirect effect can be conceptualized as an indirect-only mediation model (Zhao et al., 2010), wherein the influence of the higher-order factor on observed variables is exclusively mediated through specific factors, with no direct effect connecting the higher-order factor to the

observed variables. For clarity, in Figure 6, I denote the higher-order factor as H and the subordinate factors as S1-S3.

Figure 6

Proportionality Constraint in Higher-Order Models



In this mediation framework, the total effect of the predictor (H) on the outcomes (items) is quantified as the product $a \times b$. This product is derived by multiplying the effect of the H factor on a mediator (e.g., S1), denoted as a , by the effect of the mediator (e.g., S1) on an item, denoted as b . Notably, the a term remains constant for all items within a given subscale, implying that the differential impact of the higher-order factor on items within a specific factor is presumed to be a function of the items' loadings on their respective specific factors (e.g., $b1$ to $b4$) (Gignac, 2016; Tóth-Király et al., 2017).

Also, as depicted, a single disturbance value (ζ) is estimated for each specific factor to represent the variance in a specific latent factor that is not accounted for by the H factor. The influence of the H factor on an item was defined as the product of item loadings on the H factor (a) and the path coefficient from the specific factors to items (b). The influence of a specific

factor on an item is defined as the product of item loadings on the specific factor (b_1 - b_4) and a constant representing the relationship between the specific factor and its disturbance (c).

Consequently, the ratio of global to specific variance is equivalent to $ab/bc=ac$ for all indicators within a subscale, implying that the ratio of general factor variance to specific factor variance is consistent across all indicators within a specific factor (Chen et al., 2006).

To further elaborate, in a higher-order measurement model, the influence of the general factor is distributed uniformly across all items within a specific factor, a phenomenon referred to as *the proportionality constraint* (Reise, 2012). However, this constraint may not always hold in real-world data, as items within a factor can be influenced to varying degrees by both the higher-order factor and the specific factors.

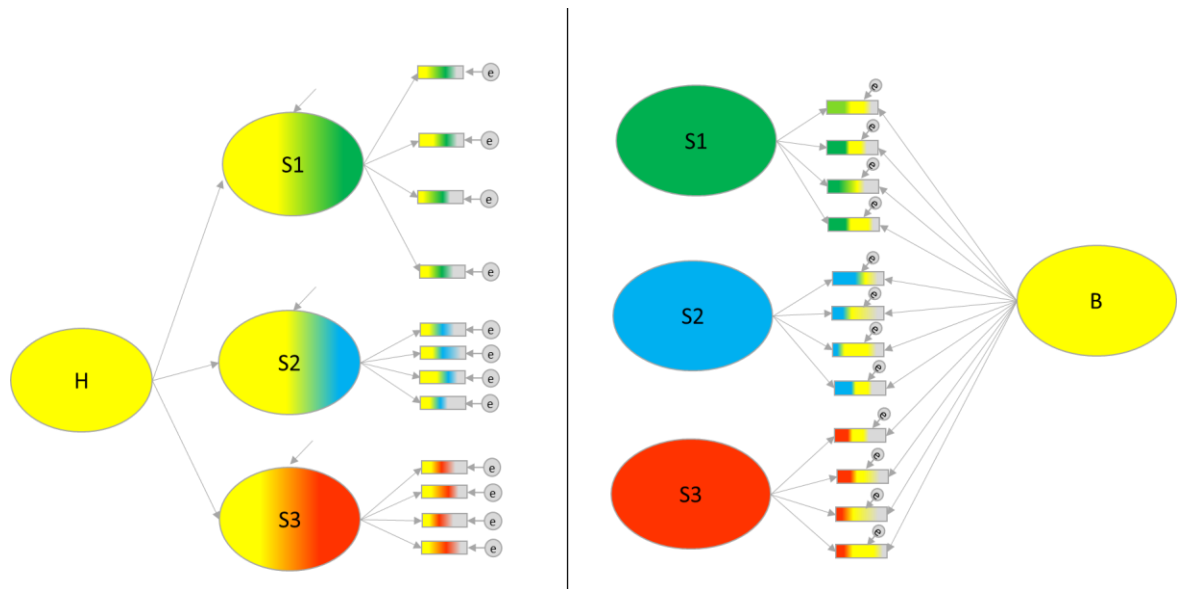
Figure 7 illustrates another important characteristic of higher-order models. In such a model, the variance of a first-order factor is composed of the variance attributable to the higher-order factor and the disturbance, which represents the shared reliable variance in the indicators that is not accounted for by the higher-order factor. Consequently, a portion of the covariance among indicators of a first-order factor is absorbed by both the higher-order and first-order factors, rendering them psychometrically redundant (Reise, 2012). To illustrate, in the context of a study, which aims to examine the criterion utility of academic challenges in predicting academic performance, the predictive validity of coexisting higher-order and first-order factors cannot be estimated due to the conceptual overlap and multicollinearity of these factors.

These limitations of the higher-order factor model serve as a rationale for considering the bifactor model as a potentially more realistic representation of the data. In a bifactor model, the *proportionality constraint* is relaxed, allowing each item to have unique loadings on the general factor and its specific factor. This feature provides greater flexibility in modeling the

relationships between items and factors (Brunner et al., 2012). In a bifactor model, the variance of each observed indicator is directly accounted for by both the specific and general factors, as well as the measurement error. Due to the orthogonality of the specific factors in a bifactor model, the general factor represents the shared variance among all observed indicators, while the specific factors represent the shared variance among items constituting a specific factor that is not already explained by the general factor (Reise, 2012). In other words, the relationships between items and general and specific factors are direct, enabling bifactor models to estimate co-existing and properly disaggregated global and specific constructs without imposing the restrictive proportionality constraints inherent in higher-order factor models.

Figure7

Disaggregation of General and Specific Factors across Higher-order and Bifactor Models



Note. The higher-order model is shown on the left and the bi-factor model on the right. The Variance of each indicator is visually decomposed into three parts: Residual errors (in Gray), the general factor (in yellow), and specific factors (in green, blue, and red). As shown, the first order and higher-order factors are psychometrically overlapping while the variance of the bifactor model is disaggregated from that of the orthogonal specific factors.

By disaggregating item variances into the variances of general and specific factors (Brunner et al., 2012; Morin et al., 2020) without multicollinearity and conceptual redundancy,

the bifactor model facilitates testing hypotheses that require operationalizing co-existing general and specific factors. An example of these hypotheses is the examination of the incremental predictive validity of coexisting specific and general factors. Incremental predictive validity refers to the additional improvement in prediction achieved by introducing a new predictor to the model. This capability renders the bifactor model a more robust tool for discerning the unique contributions of both global and specific academic challenges in predicting the outcomes.

In summary, bifactor modeling presents a more suitable alternative for representing a hierarchically organized factor structure (Gillet et al., 2020; Morin et al., 2016; Morin et al., 2020). Consequently, Morin (2023) and Morin et al. (2022) recommend avoiding higher-order models unless there are strong theoretical and/or empirical justifications for incorporating strict implicit proportionality constraints into the model.

Reliability and Dimensionality of Bifactor Models

It is crucial to note that the inclusion of a general factor in bifactor models to enhance the construct validity does not automatically enhance the instrument's reliability. Despite potential improvements in model fit, researchers must carefully examine the reliability parameters of these models (Reise et al., 2013; Rodriguez et al., 2016), as evidence suggests that traditional goodness-of-fit indices may statistically favor bifactor models (Bonifay et al., 2017, Gignac, 2008, 2016; Morgan, et al., 2015). To comprehensively evaluate the quality of a bifactor model, it is essential to employ a set of auxiliary measures (e.g., Dueber & Toland, 2023; Rodriguez et al., 2016) that facilitate the examination of the dimensionality and reliability of measurement models.

In a bifactor model, item variance and covariance can be attributed to multiple common sources, such as the general factor (g), specific factors (s), and uniqueness (random and

systematic error). Auxiliary indices provide information about the partitioning of variance among these sources, enhancing our understanding of the scale's multidimensionality (Simms, et al., 2008). By examining these indices, researchers can address questions regarding the nature of the measure's multidimensionality, the reliability of the general and specific factors (Chen et al., 2006; Raykov & Pohl, 2013; Raykov & Marcoulides, 2018), and the adequacy of total and subscale scores.

To illustrate, the variance of a specific factor in a bifactor model is the common reliable variance of its items left unexplained by the general factor. That is, in a bifactor model, specific and general factors compete in explaining the reliable item variance. Sometimes despite the multidimensional structure of a construct, the general factor accounts for a significant portion of the common variance among items remaining minimal variance for the specific factors to explain. This phenomenon, referred to as the vanishing specific factors due to a dominant general factor (Morin et al., 2020; van Zyl & ten Klooster, 2022), result in minimal and at times anomalous patterns of factor loadings on one or more specific factor.

In such a case, while the evidence from the bifactor analysis supports that the scale does embody multidimensionality, the presence of a strong general factor adequately captures the essence of the construct being assessed. The existence of such a robust general factor allows for modeling the scale as essentially unidimensional despite inherent multidimensionality. In another contrasting scenario, the specific factors can reliably account for great portion of the items' reliable variance, while the general factor represents non-intended but systematic bias, such as method effects.

The ancillary analyses of reliability and dimensionality such as explained common variance, the percentage of uncontaminated correlations, and the average relative parameter bias

(Dueber, 2017; Rodriguez et al., 2016) allow for decomposing variance across the general and specific factors and reveal these potential issues about the nature of the scale's multidimensionality. These indices also assess other aspects of the measure's psychometric characteristics. For instance, omega reliability coefficients (ω , McDonald, 2013) is used for evaluating the quality of unit-weighted total and subscale score composites, FD (factor determinacy) is used for assessing the quality of factor score estimates that can be used as proxies of the true factors, and H (construct replicability) is used to evaluate the quality of measurement models within structural equation modeling (Beauducel, 2011; Flora, 2020; Gorsuch, 2014; Grice, 2001; Hancock & Mueller, 2001; McDonald, 2013; Reise et al., 2013; Rodriguez et al., 2016; Zinbarg et al., 2005). These auxiliary indices along with the conventional criteria to evaluate them are presented in the Appendix.

Symmetric Bifactor Model Versus S-1 Bifactor Model

In the context of measurement models, reflective and formative models represent two distinct approaches to conceptualizing the relationship between latent constructs and their observed indicators. In a reflective measurement model, the latent construct is assumed to cause the observed indicators, and changes in the construct are reflected by changes in the indicators (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). The indicators are considered to be interchangeable, as they are all assumed to be random measures of the underlying construct (Jarvis et al., 2003). In contrast, in a formative measurement model, the observed indicators are assumed to cause the latent construct, and changes in the indicators lead to changes in the construct (Diamantopoulos & Winklhofer, 2001; Jarvis et al., 2003). The indicators in a formative model are not interchangeable, as each indicator captures a unique aspect of the construct and omitting an indicator may alter the meaning of the construct (Bollen & Lennox,

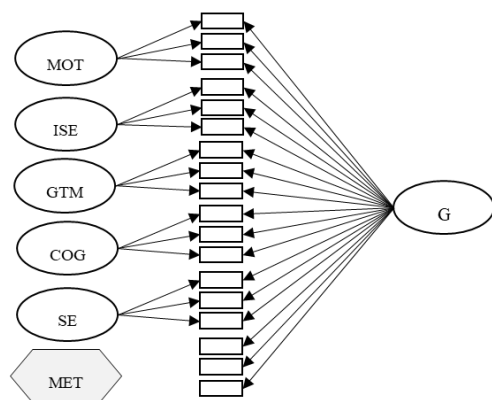
1991; Edwards & Bagozzi, 2000). The choice between reflective and formative measurement models should be based on theoretical considerations and the nature of the relationship between the construct and its indicators (Coltman et al., 2008; Jarvis et al., 2003). The measurement models discussed in this context, namely the symmetrical bifactor model and the bifactor S-1 model, are classified as reflective measurement models.

According to stochastic measurement theory, for reflective measurement models to be valid, indicators/domains should be interchangeable. Factors are considered interchangeable when they serve as random, rather than fixed, indicators of the latent constructs, implying that they can be functionally substituted for one another. In a bifactor model, this means assessing whether various specific factors reflecting a broader construct are functionally equivalent, contributing roughly equally to the overall general factor (G) without the need to functionally distinguish between these domains (Bollen & Bauldry, 2011; Bollen & Lennox, 1991; Eid et al., 2017; Eid, 2020; Jarvis et al., 2003).

In a symmetrical bifactor model (Figure 5b), all specific factors of a scale are modeled, with each item loading on both the general factor and its respective specific factor, and the general factor being defined by all items within the scale. In contrast, in a bifactor S-1 model (Figure 8), one specific factor (S) is chosen as a reference factor to directly define the general factor (G), and S-1 specific factors are estimated, while other specifications of a regular bifactor model are applied (Eid et al., 2017).

Figure 8

An example of Bifactor S-1 Measurement Model with Metacognitive Challenges as the Reference domain



The bifactor S-1 model offers advantages over the symmetrical bifactor model. It is an alternative to the symmetrical bifactor model when factors are expected to contribute asymmetrically to the general factor. In a symmetrical bifactor model, the meaning of the G-factor can become ambiguous when specific factors differentially contribute to the general factor and are not functionally equivalent. The G-factor in a symmetrical bifactor model represents shared variance among all indicators, while S-factors capture unique variance shared among indicators of a subscale beyond what is explained by the G-factor (Bureau et al., 2023). When specific factors differentially contribute to the general factor and are not functionally equivalent, the meaning of the G-factor can become ambiguous, requiring study-by-study interpretation based on factor loadings. The shared variance among all indicators in the G-factor may also be contaminated by construct irrelevant source of variance (e.g., common method bias).

In contrast, in an S-1 bifactor model, by choosing the reference domain, the G-factor primarily derives its meaning from items within this domain, with decreasing influence from items in other domains as indicated by their relative factor loadings on the G-factor. This leads to factors with clearer conceptual boundaries and stability, as the G-factor is primarily defined by the reference factor items, making it less likely to fluctuate across studies compared to a data-driven G-factor in symmetrical bifactor models (Haywood et al., 2021; Heinrich et al., 2020;

2023). Also, interpretation of the G-factor in an S-1 model remains consistent when other specific factors are added or removed, as long as the reference domain remains the same, facilitating efficient knowledge accumulation in systematic reviews and meta-analyses.

Also, the bifactor S-1 model is less prone to anomalous results in empirical applications, as it can be defined as a reformulation of first-order correlated factor models with the same levels of global fit, which generally have fewer estimation problems (Eid et al., 2018). However, choosing the reference factor should be based on theoretical considerations, not statistical evidence (Burns et al., 2020; Eid, 2020; Eid et al., 2017; Heinrich et al., 2020).

Additionally, specific factors in S-1 models represent the part of the specific factor that cannot be explained by the reference factor, providing a clearer interpretation than in symmetrical bifactor models. In S-1 models, correlations between S-factors can be estimated and interpreted, representing what two S-factors have in common once the effect of the reference factor is accounted for (Bureau et al., 2023; Eid et al., 2017). For example, if metacognitive challenges are set as the reference domain in the context of an academic challenges scale, the motivation scale within the bifactor S-1 model specifically measures the part of the variance in motivational challenges that is independent of metacognitive challenges influences.

To summarize, when dealing with constructs that have specific factors with non-equivalent functionality, defining the general factor *a priori* is a more valid approach than interpreting it *post-hoc*. In such cases, the bifactor S-1 model, which anchors the general factor on a predetermined reference domain, offers advantages over the symmetrical bifactor model.

Integrating Cross-Loadings Within the Pre-Specified Measurement Model

Aside from not modeling a potential general factor underlying the responses, correlated CFA models fail to account for another construct-relevant psychometric characteristic. Also

known as Independent Cluster CFA (ICM-CFA), correlated CFA models assume local independence of clusters. This ICM assumption postulates that each observation has only one construct-relevant source of true variance. Put simply, each item loads exclusively on its target factor, and the associations between the item and all non-target factors, known as cross-loadings, are constrained to zero (Xiao et al., 2019). For example, in an ICM CFA (i.e., regular correlated CFA), an item from the motivation challenge subscale is assumed to be caused solely by the motivation challenges latent factor, and the association between the item and other factors (e.g., metacognitive challenges) is forced to be zero.

The ICM-CFA assumption of zero cross-loadings can be problematic when dealing with multidimensional constructs (Morin, 2023; Morin et al., 2016, Morin et al., 2020). It is often unlikely and ecologically invalid to assume that items on a subscale are not influenced by non-target factors, particularly when an instrument measures conceptually related constructs like academic challenges (see Marsh et al., 2014; Morin et al., 2016; Morin et al., 2020). In the case of the academic challenges scale, items are likely to relate to factors other than the primary ones to which they are assumed to load *a priori*. For instance, the item “Putting things off (procrastinating)” from the Initiating & Sustaining Engagement Challenges scale (ISE) may cross-load with items from the Cognitive Challenges scale, as struggling with comprehension can lead to procrastination. It may also cross-load with the item “Believing I can do my work” from the Motivation Challenges scale, as low self-efficacy can contribute to procrastination. Similarly, cross-loadings may occur with the item “Knowing how to tackle tasks” from the Metacognitive Challenges scale, as not knowing how to approach tasks can also lead to procrastination. The item may also cross-load with “Meeting deadlines” from the Goal & Time Management Challenges scale, as procrastination directly affects meeting deadlines. Finally, it

may cross-load with “Managing my emotions/feelings” from the Social & Emotional Challenges scale, as emotional difficulties can contribute to or be an outcome of procrastination, highlighting the overlap between procrastination and emotional challenges.

Forcing cross-loadings to zero in correlated CFA models can lead to artificially inflated factor correlations or overestimated item loadings on the general factor (Murray & Johnson, 2013). This is because the unmodeled shared variance between items of each scale and items of the other factors is pushed into factor correlations, resulting in distorted estimations of relationships among the factors and between the items and the general factor. These biases in the measurement model can potentially bias the structural path estimates when investigating the association of the measurement model with other variables (Asparouhov & Muthén, 2009; Marsh et al., 2014; Morin et al., 2020; Shao et al., 2022; Zhang et al., 2023).

Additionally, statistical simulation studies have shown that ESEM generally yields less biased parameter estimates compared to ICM-CFA. To elaborate, while failing to model even small cross-loadings results in biased estimates of factor correlations and regression coefficients, allowing the free estimation of minor cross-loadings still produces unbiased parameter estimates (Asparouhov et al., 2015; Mai et al., 2018; Wei et al., 2022).

Furthermore, the restrictive nature of ICM-CFA frequently leads to models with suboptimal levels of fit, even though they may represent replicable factor structures (e.g., Marsh et al., 2013; Marsh et al., 2009). As a result, researchers often attempt to improve model fit through *post hoc*, data-driven procedures guided by misspecification indices, such as modification indices. One common practice is to add covariances between item residuals, a modification that often lacks theoretical grounding and may lead to measurement structures that are not replicable (Landis et al., 2009; Hermida, 2015). Correlating measurement errors typically

suggests the presence of an overlooked latent factor, or it may stem from item sequencing or similar phrasing (Cole et al, 2007; Brown, 2015). When researchers cannot clearly identify the reason for the correlation between two items beyond their correlation caused by their target factor, the modifications made to the measurement model may merely serve to artificially improve fit indices rather than genuinely enhance the stability and replicability of the measurement model.

Exploratory Structural Equation Modeling (ESEM) has been proposed as an alternative analytical method to address these limitations of ICM-CFA models. ESEM allows items to primarily load on their target factors while also permitting item loadings on non-target factors (i.e., cross-loadings). This means that while each item primarily loads on its target factor, contributions from other non-target factors in determining the variance of the item are also allowed.

ESEM is described as a synthesis of Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), designed to overcome the limitations of both EFA and CFA approaches (Asparouhov & Muthén, 2009). EFA involves data-driven statistical techniques used to identify the minimal number of relevant dimensions required to explain the covariation among a set of observations. In EFA, cross-loadings are freely estimated (Costello & Osborne, 2005; Fabrigar et al., 1999). In contrast, CFA is a theory-driven approach aimed at testing the alignment between a theoretical model and the data, where cross-loadings are constrained to zero (Brown, 2015, Kline, 2015). While EFA does not allow for the investigation of the global fit of the model, since factors are derived in a data-driven way and labeled post hoc, CFA's restrictive assumption regarding cross-loadings makes it vulnerable to inflated factor correlations and deflated goodness of fit, threatening the psychometric validity of the measurement model. ESEM

combines the benefits of these two approaches enabling the assessment of global fit of an *a priori* model while accommodating complex relationships among items and factors in real-world data (Asparouhov & Muthén, 2009; Marsh et al., 2014).

There are variations of ESEM that range from more exploratory to confirmatory approaches (see Prokofeva et al., 2023, for a concise review). These variations can be grouped into two pathways. The first pathway addresses the factor structure primarily in a confirmatory manner, freely estimating item loadings on their primary allocated factors while uniformly constraining non-primary loadings/cross-loadings to near-zero values using calculation thresholds (Marsh et al., 2014; Morin et al., 2020). Conversely, the second pathway takes a more exploratory approach to the factor structure. It begins by conducting factor-specific EFAs to estimate item loadings on their primary allocated factors. Subsequently, the factor-specific EFAs are expanded to include non-primary items, allowing for the estimation of cross-loadings, which are addressed using thresholds derived from the EFA procedures (Silvestrin, 2024; Prokofeva et al., 2023).

As the factorial structure of academic challenges was known *a priori*, the version employed in the studies comprising this dissertation leans fundamentally towards the confirmatory approach. This method involves the *a priori* specification of the scale's factorial structure, allowing for the free estimation of item loadings on their designated factors. Then rotation techniques such as geomin or oblimin for first-order models (Giordano & Waller, 2020) and target rotation for the bifactor models are used to minimize cross-loadings on non-target factors, aiming to keep these as close to zero as possible while still being freely estimated (Marsh et al., 2014; Morin et al., 2020).

In sum, ESEM emerges as an analytical framework that bridges the gap between exploratory and confirmatory factor analyses. ESEM is particularly valuable in cases where strict zero cross-loadings imposed by traditional CFA may be overly restrictive, leading to inflated factor correlations and poor model fit. By incorporating both the exploratory nature of EFA and the hypothesis-testing capabilities of CFA, ESEM enables researchers to examine complex factor structures more accurately, especially in multidimensional constructs where some degree of construct overlap is theoretically plausible. These features empower ESEM to provide a more realistic and ecologically valid representation of these interconnected multidimensional constructs, as it accommodates the complex relationships that often exist among items and factors in real-world data (Asparouhov & Muthén, 2009; Marsh et al., 2014).

Sequential Strategy for Testing Alternative Measurement Models

Two key characteristics relevant to multidimensional constructs were discussed. First, the potential existence of a general factor underlying item responses was examined. Second, the possibility of integrating cross-loadings within the measurement model, rather than assuming factors as independent clusters, was explored.

To address these aspects, bifactor model was identified as a suitable approach for modeling the first characteristic, and ESEM as an analytical method for the second. Notably, these models can be combined into a unified framework, known as bifactor-ESEM, to address both objectives simultaneously.

The bifactor-ESEM model can be viewed from two perspectives. On one hand, it can be seen as extending the bifactor CFA by incorporating cross-loadings. On the other hand, it can be viewed as augmenting the first-order ESEM by introducing a general factor. In a bifactor-ESEM model, the variance of each item is conceptualized as being derived from four distinct sources:

1. Variance attributable to the target specific factor
2. Variance attributable to the general factor
3. Variance attributable to non-target factors
4. Variance unique to the item (residual variance)

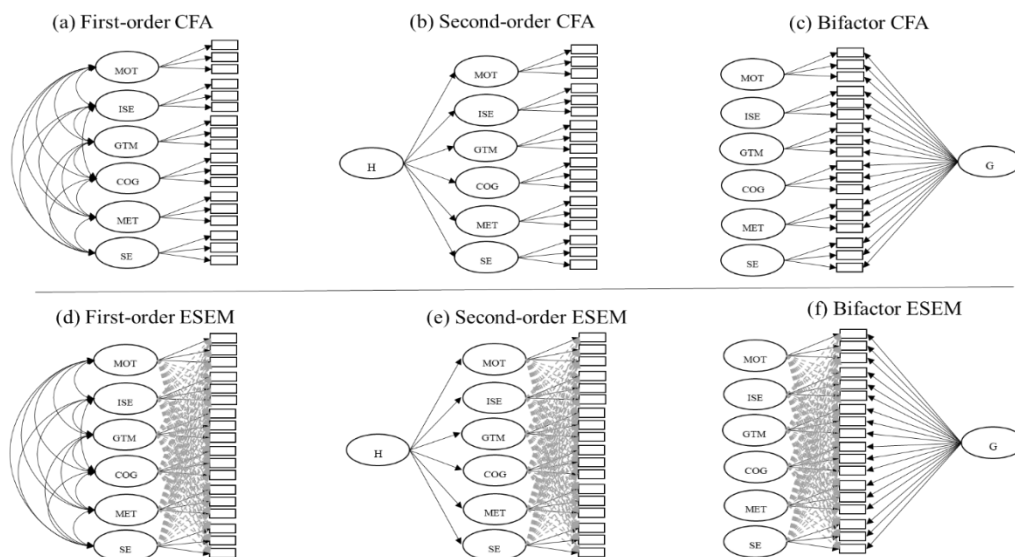
The bifactor-ESEM model is distinguished from the bifactor-CFA by the third source of item variance, while it is set apart from a first-order ESEM by the second source. This integrated approach allows for the modeling of a potential general factor while the strict assumption of local independence is relaxed through the estimation of cross-loadings.

The selection of an appropriate measurement modeling approach is a critical decision for researchers. This choice should be primarily guided by theoretical considerations, followed by empirical evidence supporting the better fit of a given measurement model. However, practicality and parsimony are also important factors to consider.

To determine the most suitable model, researchers may need to evaluate how incorporating cross-loadings or adding a general factor improves their model's fit. This evaluation helps in deciding whether a more complex model is warranted or if parsimony should be prioritized. Addressing these questions requires a systematic approach. A recommended strategy involves the sequential testing of alternative measurement models (Morin, 2023; Morin et al., 2016; Morin et al., 2020). As shown in Figure 9, this process typically includes examining (1) a unidimensional CFA, (2) first-order CFA, (3) second-order CFA, (4) Bifactor-CFA, (5) First-order ESEM, (6) Second-order ESEM, and (7) Bifactor-ESEM.

Figure 9

Correlated, Second order, and Bifactor CFA and ESEM Models for the Academic Challenges Scale



Note. Unidimensional model is not included.

Abbreviations. H: Higher order; G: General; MOT: Motivation, ISE: Initiating & Sustaining Engagement; GTM: Goal & Time Management; COG: Cognitive, MET: Metacognitive Challenges; SE: Social & Emotional Challenges

Testing and reporting unidimensional and first-order correlated CFA (9a) models as initial steps provides valuable context for better understanding the findings from more advanced modeling approaches (Eid, 2020). Higher-order (9b) and bifactor (9c) models are employed to explore the potential existence of an overarching latent construct underpinning the observations. As mentioned, these measurement models (correlated factor, higher-order, and bifactor CFA) should be tested within the ESEM framework. In the case of correlated ESEM (9d), an item's variance is explained by multiple latent factors simultaneously, without necessarily implying the presence of an unmeasured factor, such as a general factor. The ESEM variants of higher-order (9e) and bifactor (9f) models are tested to complement the objective of modeling the general factor while incorporating cross-loadings. The assessment of the relative strengths and weaknesses of measurement models within this sequential approach involves a comprehensive, multistep evaluation process that extends beyond simply examining model fit. A concise

overview of the key evaluative criteria employed in this process can be found in the Study 3.

Table 5 provides a summary of the pros and cons of the modeling approaches discussed so far.

Table 5

Advantages and Disadvantages of Analytic Methods Discussed

<i>Measurement Models</i>	<i>Pros</i>	<i>Cons</i>
CFA (Figure 9, models <i>a, b, c</i>)	Parsimoniousness	Forcing cross-loadings to zero ¹
ESEM (Figure 9, models <i>d, e, f</i>)	◦ Allows for specification of theoretical models ◦ Allows for modeling of cross loadings	◦ Less parsimony ◦ Convergence Problems

Advantages and Disadvantages of Second-order versus Bifactor Measurement Models

Second-order (Figure 9, models <i>b & e</i>)	Parsimoniousness	◦ the ratio of H factor variance to specific factor variance is the same for all indicators within a specific factor² ◦ First order and second order factors are psychometrically redundant
Bifactor (Figure 9, models <i>c & f</i>)	◦ Disaggregated g and s factors ◦ Allowing for unequal contribution of g to the items	◦ Less parsimony ◦ Vanishing specific factors when the general factor explains a large proportion of the variance

Advantages and Disadvantages of Symmetrical Bifactor versus S-1 Bifactor Measurement Models

Symmetrical Bifactor	◦ Higher Fit to the data ◦ Operationalizing all specific factors	◦ More Statistical Anomalies ◦ Model identifications Problems ◦ Non-replicable G factor when S factors are not functionally equivalent
S-1 Bifactor	◦ Conceptual Clarity and Stability of the G ◦ Less statistical Anomalies ◦ Appropriate when S factors are functionally non-equivalent	◦ Vulnerability to arbitrarily selecting the G factor without a theoretical basis ◦ Lower fit

Note.

¹Restrictive assumption of independence of clusters

²Proportionality Constraints

Evaluating Criterion Utility of the Measurement Model

The associations of factors with criterion variables can also be evaluated across different measurement models. For example, in the context of academic challenges, researchers might

examine how different models of challenge structure (e.g., first-order, second-order, or bifactor) relate to outcomes such as academic performance, student engagement, or well-being. This evaluation provides valuable insights into how the choice of analytical model affects the interpretation of relationships between constructs. For instance, in a first-order CFA or ESEM, the associations between each factor and criterion variables are directly observable while in a second-order model, the relationships between lower-order factors and criterion variables are mediated through the higher-order factor, potentially altering the interpretation of these associations. In a bifactor model (CFA or ESEM), the general factor and specific factors may show different patterns of associations with criterion variables. This can reveal how differentially the general factor and specific factors are related to certain outcomes. As an example, the inclusion of a general factor in bifactor models might reduce the associations of specific factors with criterion variables, as some of the shared variance is now accounted for by the general factor. Comparing ESEM to CFA models can show how allowing for cross-loadings affects the strength and pattern of associations with criterion variables.

In sum, by systematically comparing these models, researchers can assess the relative improvements in fit and determine the most appropriate balance between model complexity and parsimony. The sequential approach discussed allows for a comprehensive evaluation of different factor structures and their implications for understanding the construct under study. These evaluations can also help researchers determine which model provides the most meaningful and interpretable relationships with external criteria. This insight enables researchers to make more informed decisions about which measurement structure best represents the construct under study, both from a theoretical standpoint and in relation to important outcome variables.

Zooming In: Moving from Scale Multidimensional Structure to Item-Level Precision

An item-level analysis can serve as a complement to the analysis of scale structure, providing a more nuanced and comprehensive understanding of the measurement instrument. While scale-level analyses, such as factor analysis, offer valuable insights into the overall structure and dimensionality of a measure, they may obscure important item-specific characteristics and behaviors. Item-level analyses, particularly through Item Response Theory (IRT), allow researchers to examine the performance of individual items across different levels of the latent trait, revealing subtle variations in item difficulty, discrimination, and information that are not readily apparent in scale-level analyses. This granular approach enables the identification of potentially problematic items, assessment of the appropriateness of response categories, and evaluation of measurement precision across the spectrum of the latent trait. Furthermore, item-level analyses can uncover differential item functioning across subgroups, ensuring fairness and psychometric validity of individual items across diverse populations.

By combining scale-level and item-level analyses, researchers can develop a more robust and refined measurement instrument, optimizing its psychometric properties and enhancing its utility in both research and applied settings. This dual-level approach ensures that the scale not only exhibits a sound overall structure but also comprises items that function effectively to measure the intended construct with precision.

A fundamental distinction between Item Response Theory (IRT) and scale-level analyses such as CFA lies in their analytical focus. CFA offers a global perspective, evaluating how well a hypothesized model aligns with empirical data. In contrast, IRT provides a more detailed, item-level analysis. While similar objectives could be achieved by examining local parameter estimates in a CFA (such as item-factor loadings, residual errors, and model-based internal

consistency indices), the primary goal of CFA remains the evaluation of the global fit of a theoretical model to the data.

In the subsequent sections of this chapter, I will first introduce the details of Item Response Theory and then discuss the added value of employing IRT to complement the scale-level analyses discussed earlier. This comparison will highlight how IRT's granular item-level focus extends and refines the insights about the multidimensional structure gained from those analyses.

Item Response Theory (IRT)

Item Response Theory (IRT) is a psychometric method for evaluating items and the quality of response scales. At its core, IRT evaluates individual items on a scale and aggregates this information to assess the overall properties of the scale. The foundational principle of IRT is that the likelihood of a particular response, such as selecting an option on a Likert scale, depends on an underlying latent attribute (De Ayala, 2013). This latent dimension, often referred to as *theta* (θ), is a person parameter that indicates the extent to which an individual possesses a specific latent attribute. For instance, in the context of academic challenges, this parameter would measure a student's perceived level of a particular challenge. *Theta* is typically standardized to follow a standard normal distribution with an average of zero. In IRT, the focus is on modeling the relationship between θ as a person parameter and the responses to items. Put differently, IRT is an item-level assessment of the measurement precision of items across the range of the latent attribute (θ). In addition to the person parameter, there are also key item

parameters in IRT, namely discrimination (α) and threshold¹ (β) parameters. These parameters provide insights into how effectively each item on a scale measures the latent attribute. The term discrimination describes the ability of an item to differentiate between individuals who have different levels of the underlying attribute that the scale is intended to measure. On the other hand, thresholds refer to the levels of the latent attribute above which individuals are likely to endorse higher response categories (Embretson, & Reise, 2013).

Through IRT, it is possible to identify items that offer less precise measurement and evaluate the reliability of each item in measuring the latent dimensions. The concept of information in IRT is a function of both discrimination and threshold parameters. Examining information parameters allows for estimating how much can be learned about a person's level on the latent attribute from their response to an item or a set of items within a scale. Such an examination provides information about the degree to which items measure the intended range of the latent attribute and into the specific levels of latent attributes where they provide suboptimal measurement (Baker, 2001).

Graded Response Model

The self-reported academic challenge is measured using a 5-point ordinal Likert scale, therefore, a polytomous variant of IRT is needed for evaluating its psychometric properties. An example of such polytomous IRT models is Samejima's Graded Response Model (GRM; 1997, 2016). GRM designed for analyzing items with ordered response categories, such as Likert scales. GRM estimates the two discussed types of parameters for each item (i.e., discrimination and threshold). The discrimination parameter in GRM (α) indicates how well an item

¹ The term *difficulty* is commonly used to denote threshold in assessment contexts. However, I prefer the term *threshold* as it more accurately reflects measures evaluating non-ability constructs.

differentiates between individuals with varying levels of the latent attribute. α can be interpreted as the strength of association between an item and the latent dimension, somewhat akin to factor loading (De Ayala, 2013). Items with low discrimination are problematic because they imply that the likelihood of choosing a higher response option does not significantly increase as the latent attribute increases, which is undesired in most assessment contexts. Therefore, such items might not effectively measure the intended latent attribute and could potentially distort the overall scale's reliability.

Discrimination parameters in a GRM model can be constrained or unconstrained. In constrained models, the discrimination parameters across all items within a scale are set to be equal. Conversely, in unconstrained models, these discrimination parameters are freely estimated.

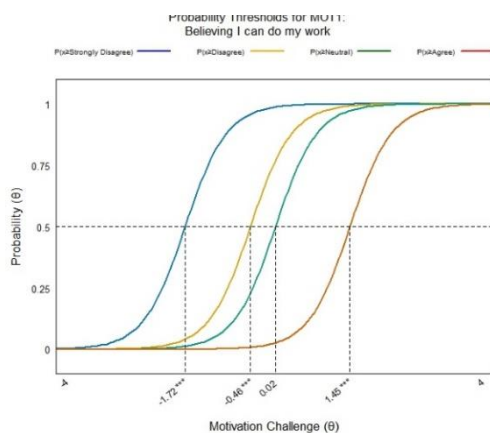
Also, in the GRM, there are $k - 1$ threshold parameters for an item with k response categories. For example, in a 5-point Likert scale, there would be 4 threshold parameters (β_1 , β_2 , β_3 , and β_4) for each item. In the GRM, the threshold parameters are cumulative, which means that they represent the probability of selecting a particular response category or any higher category, given a specific level of the latent attribute. Specifically, GRM threshold parameters represent the level of the latent attribute at which an individual has a 50% probability of selecting a particular response category or higher. This parameter can vary across a wide range, typically falling between zero to +2.5 (Baker & Kim, 2017). In the context of an academic challenges scale, higher threshold parameters indicate that a higher level of the latent attribute (i.e., more severe academic challenges) is required for an individual to endorse higher response categories, while lower threshold parameters suggest that even individuals with lower levels of academic challenges are likely to select higher response categories for that item.

Graphs Depicting IRT Parameters

One of the key advantages of using IRT, particularly unidimensional IRT, is its ability to visually represent the quality of measurement in terms of the relationship between item and person parameters. To this end, Boundary Characteristic Curves (BCCs, Figure 10) and Item Response Category Characteristic Curves (ICCs, Figure, 11) provide a graphical representation of how the probability of endorsing a particular response category changes across different levels of the latent attribute (Ostini & Nering, 2006). In addition, Item Information Curves (IICs) for each item (Figure 12a) and for the whole subscale (Figure 12b) show the amount of information that each item (or the subscale as a whole) provides at different levels of the latent attribute (Van Der Ark, 2001). IICs can be converted to reliability graphs indicating the reliability distribution for items and the subscales across the range of the latent attribute (Figure 13). The Expected Total Score plot (Figure 14) helps convert the latent attribute scores into the total score in the original observed units of the scale. Examples of these graphs and brief captions describing how to interpret them are provided below.

Figure 10

Boundary Characteristic Curves (BCCs) for an item of the Motivation Challenges Scale

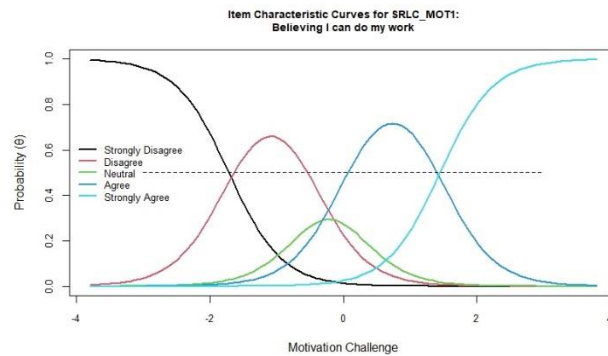


Note. The boundary characteristic curves (BCCs) for the item "Believing I can do my work" offer a visual representation of the probability thresholds' distances between response categories and the likelihood of endorsing each response category at different levels of the latent attribute (Motivation Challenge).

- The color-coded curves represent response categories (e.g., Strongly Disagree, Disagree, Neutral, Agree).
- The vertical axis (y-axis) denotes the probability of endorsing each response category.
- The horizontal axis (x-axis) represents the levels of the latent attribute (Motivation Challenge).
- The numbers on the x-axis indicate the threshold points where the probability of endorsing the next category becomes higher.

Figure 11

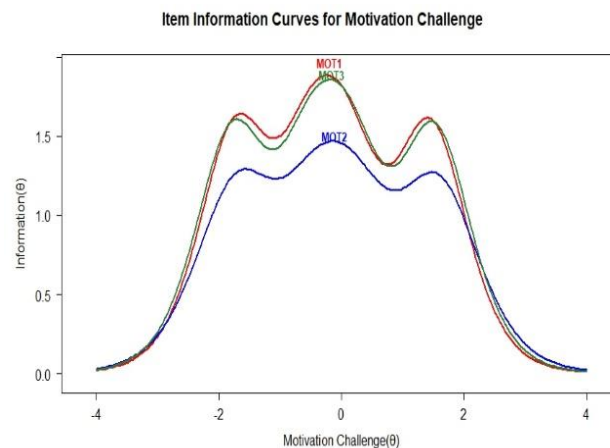
Category Characteristic Curves (CCCs) for an item of the Motivation Challenges Scale



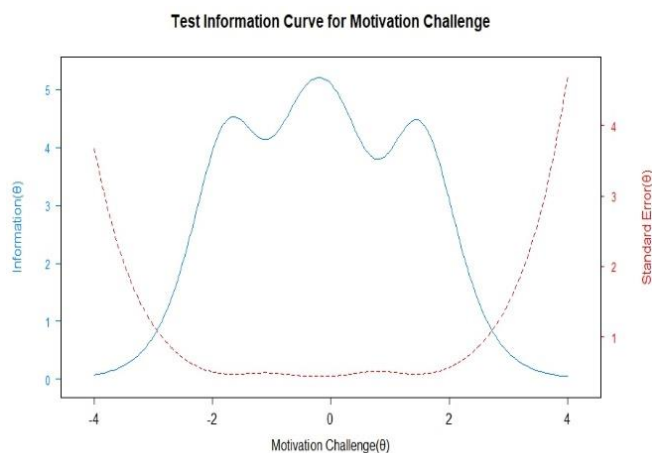
Note. Each curve illustrates the probability of choosing a specific response category (e.g., Strongly Disagree, Disagree, etc.) across the levels of Motivation Challenge for an item. The x-axis denotes the latent attribute, while the y-axis represents the probability. The points at which two adjacent curves intersect indicate the thresholds where the transition from one response option to the next is equally probable.

Figure 12

Item Information Function (IIF) Graphs (a) and Test Information Curve (TIC) and Standard Error Graphs (b)

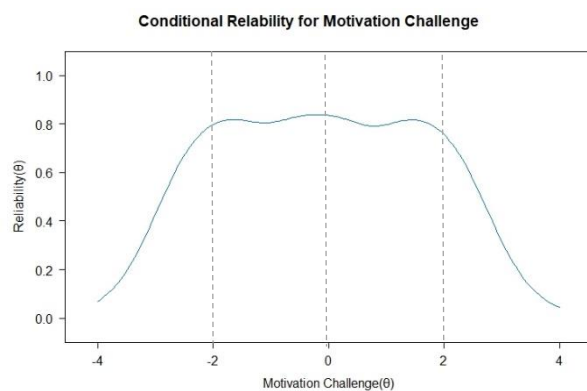


Note. The item information function (IIF, 12a) displays the item information curves for the Motivation Challenges items. Each curve represents the amount of information an item provides about the latent attribute (Motivation Challenge) at different levels. The IIF helps identify the levels of the latent attribute at which each item is most informative and is valuable for selecting items that offer the most information at various levels of the latent attribute.

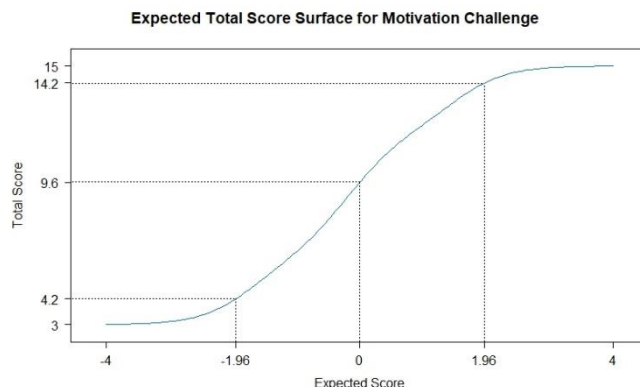


Note. The test information curve (TIC, 12b) aggregates the information from all items in the scale and presents the overall informativeness of the Motivation Challenges Scale across various levels of the latent attribute, while the standard error curve, which is inversely related to the information curve, illustrates the measurement imprecision. Higher test information and lower standard error values indicate higher measurement precision.

Figure 13
Conditional Reliability Graph



Note. This graph illustrates the conditional reliability of the Motivation Challenges Scale, depicting how reliability fluctuates across various levels of the latent attribute. It aids in identifying ranges of the latent attribute where the scale may require enhancement to improve reliability by showcasing the scale's reliability at different levels of the latent attribute. Higher reliability values signify more consistent and dependable measurement. Dotted vertical lines at specific attribute levels ($\theta = -2, 0, 2$) serve as reference points for interpreting reliability at crucial points along the latent continuum.

Figure 14*Expected Total Score Graph*

Note. This graph illustrates the relationship between the latent attribute (Motivation Challenge) and the anticipated overall score on the Motivation Challenges Scale. It facilitates the interpretation of various latent scores by expressing them in terms of the total scores using the scale's original metrics.

Assumptions of IRT

Several fundamental assumptions must be satisfied in classic IRT, for the model to be valid and reliable. These include unidimensionality, local independence, monotonicity, and differential item functioning (Baker & Kim, 2017).

Unidimensionality. Assumption of unidimensionality posits that all items on a scale should measure a single latent dimension. In other words, a single dimension should be sufficient to explain the response patterns to the items. This means that the scale or test is measuring one, and only one, underlying attribute. If a test is truly unidimensional, a person's response to any one item should be predictable based on their responses to the other items. If a test or scale is not unidimensional, the estimates of item parameters and person attributes may be biased or inaccurate, potentially leading to misleading conclusions about the measured trait, ability, or attribute. This assumption can be assessed using factor analysis, where ideally, all items should load significantly on a single factor (Ostini & Nering, 2006).

Local Independence. The assumption of local independence implies that the latent attribute being measured should account for a substantial portion of the variance in responses to the items within a scale. To test for local independence, the residual correlations between items are examined. For a scale to be locally independent, these residual correlations should ideally be weak and not statistically significant, indicating that once the latent attribute is accounted for, the responses to different items within a scale are unrelated to each other (Edwards, et al., 2018).

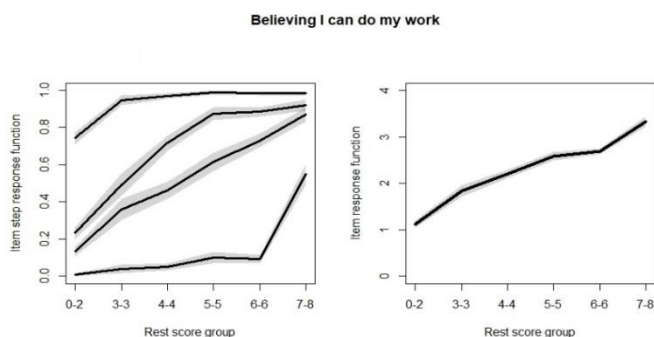
Monotonicity. Monotonicity in IRT refers to an important assumption about the relationship between the estimated latent dimension being measured (e.g., trait, ability, or attribute) and the probability of a given response to an item. Specifically, monotonicity means that as the level of the latent attribute increases, the probability of selecting a higher category response (or demonstrating a higher level of ability/performance on an item) should also increase, or at least not decrease (Kang, et al., 2018). In IRT analysis, checking for monotonicity is an important step in ensuring that the items on a scale or test are functioning properly and are appropriate for measuring the intended latent dimension. Violation of this assumption indicates a problem with the item itself or with the way the latent dimension is being conceptualized or operationalized. An item that does not exhibit monotonicity might be confusing, poorly worded, or not aligned properly with the underlying construct, leading to responses that do not reflect the intended measurement of the latent attribute.

Monotonicity plots (e.g., Figure 15) can be used to visually assess whether the monotonicity assumption is met (Mokken, 1971; Sijtsma & van der Ark, 2017). In these plots, the x-axis represents the overall sum score on the scale, excluding the score from the item under investigation, while the y-axis represents the item response function (IRF) for individual items.

These plots visualize the probability of a particular response to an item against the level of the latent attribute.

Figure 15

Monotonicity Plot for an Exemplar Item



Note. This plot displays two plots for an example item. The plot on the left shows the monotonicity plot, which includes four item step response functions, each representing the cumulative probability of responding at or above a specific category threshold. The plot on the right combines the four ISRFs into a single curve, showing the cumulative probability of responding at or above each category threshold as a function of the latent trait level.

The individual item step response function plots can be aggregated to represent a cumulative item response function for an item. Ideally, the IRFs for all items should show a relatively diagonal line, indicating a direct and consistent relationship between the scale's total score and the item responses. To meet the monotonicity assumption, there should be an absence of any downward dips in the IRFs. Such dips would suggest that higher levels of the latent attribute are associated with lower responses on the ordered Likert scale, which is counterintuitive in this context.

Differential Item Functioning. Assumption of Differential Item Functioning (DIF) ensures that the item response patterns are consistent across different demographic groups, such as gender. The absence of DIF suggests that the items function similarly across these groups,

implying that the scale is fair and unbiased in measuring the latent attribute across the diversity of population it intends to represent (Chalmers et al., 2016).

Each of these assumptions plays a crucial role in ensuring the integrity and applicability of an IRT model, making them critical requirements in the development and evaluation of psychometric scales.

Unidimensional versus Multidimensional Item-level Analyses

While unidimensional IRT models, such as the Graded Response Model (Samejima, 1997, 2016), assume that a single latent dimension is sufficient to explain the response patterns to the items, multidimensional IRT models (MIRT; Reckase, 2009) allow for the possibility of multiple latent dimensions influencing item responses. The choice between unidimensional and multidimensional IRT analyses depends on the complexity of the construct being measured and the structure of the scale (Embretson & Reise, 2000). Unidimensional IRT models are more parsimonious and easier to interpret, but they may oversimplify the underlying structure of the data if multiple dimensions are present. Conversely, multidimensional IRT models can capture the nuances of multidimensional constructs but may be more complex to estimate and interpret (de Ayala, 2009). Moreover, multidimensional IRT models generally require larger sample sizes compared to unidimensional models, as they estimate a greater number of parameters (Reise et al., 2007). An example of a multidimensional IRT model is the bifactor polytomous multidimensional Item Response Theory, which can provide insights into the general and specific aspects of the measure (see Immekus, et al., 2019; Toland, et al., 2017). However, unlike bifactor factorial models, it cannot be implemented within ESEM framework to offer greater exploratory flexibility in explicit modeling of cross-loadings.

Consequently, to complement the prior scale-level analyses of multidimensionality, a unidimensional IRT approach appears to be a more suitable analytic strategy. Conducting a multidimensional IRT analysis in addition to a multidimensional factor analysis may not yield substantial additional information, as both methods aim to capture the multidimensional nature of the construct. The remainder of this section discusses why a unidimensional IRT approach is a better complement to the prior scale-level analyses of multidimensionality. I will explore the advantages of evaluating the academic challenges scales as separate unidimensional constructs and then as a single general dimension. This strategy provides a more comprehensive understanding of the scales' psychometric properties and their potential for independent use in various contexts.

Evaluating the Adequacy of Composite Scores for the Scale Practical Use

The academic challenges scale has been developed to serve both research and practical purposes in educational settings. In practical settings such as institutions, classrooms, or online digital reports where the scale is used for diagnostic screening or self-assessment, latent measurement modeling may not be feasible or accessible. Additionally, factor indeterminacy presents a challenge for computing factor scores for covariance-based structural equation models (Rigdon, 2019). As a result, non-refined composite scores such as sum or average scores (DiStefano et al., 2019) are often used as substitutes for latent factors. For example, in classrooms, educators often calculate sum or average scores, and in digital reports, sum/average scale scores are reported back to respondents in tabular or visual forms. Moreover, researchers may be interested in using a regression-based approach to conduct conditional process analysis (Hayes & Rockwood, 2020), for which computing composite scores for the scales is necessary.

This approach, however, applies restrictive assumptions on modeling measurement error, such as assigning equal weights to contributing items in calculating average scores or disproportionately allocating greater weights to scales with more items when computing overall sum scores. It is important to note that the validity of interpretations of a model derived from an analytic procedure such as CFA, which explicitly accounts for measurement error, cannot be directly used as an indication of the validity of interpretations of composite scores (McNeish & Wolf, 2020). Certain conditions, such as unidimensionality and high internal consistency of factors, must be met for composite scores to yield relatively consistent results with covariance-based estimates of factor analysis and structural equation modeling (Widaman & Revelle, 2023).

Analyses such as unidimensional IRT and Classic Test Theory (CTT) are helpful in providing insights into the extent to which composite scores deviate from or align with true factors. Moreover, factor scores derived from IRT can help assess the quality of these scores. For example, EAPsum (Expected A Posteriori Sum) employs a concurrent Bayesian calibration method that transforms summed scores into scaled scores. This approach takes into account item difficulty and discrimination, effectively integrating the IRT-based properties of the scale with the practicality of summed scores (Thissen et al., 1995). Unidimensional IRT and CTT analyses collectively provide essential insights into the appropriateness of using composite scores across various research and practical contexts and help ensure that the conclusions drawn from such scores are valid and reliable.

Evaluating the Psychometric Adequacy of Each Subscale as a Discrete Scale for Research Applications

Structural Equation Modeling (SEM) has gained widespread popularity as an analytic framework in social sciences and psychology (Karimi & Meyer, 2014; Tarka, 2017). However,

to avoid overparameterization, convergence issues, and preserve statistical power, it may not always be feasible to include all subscales of a scale in the model, especially when the scale consists of multiple factors and the research questions do not directly address all those specific factors. As a result, researchers often include and estimate a single factor or a subset of factors in their analyses. Since most multidimensional constructs are validated using a correlated factor CFA, information about the global fit of such models is frequently taken as evidence of the construct validity of the selected scale or subset of scales. Mathematically, a multidimensional measurement model comprising k factors is equivalent to k unidimensional measurement models that are allowed to covary within a structural model. As the concept of fit represents the agreement between the covariance structure of the data and the one reproduced by the specified model, the overall psychometric adequacy of a multidimensional latent model does not necessarily guarantee the local adequacy of all its subscales as independent or subset measures. Unidimensional IRT allows for a detailed examination of the psychometric properties of individual subscales as standalone constructs and independent measures.

Additionally, unidimensional IRT offers the advantage of visual representation through intuitive graphs, which effectively illustrate measurement quality and highlight problematic items. These visual aids enhance communication of results more effectively than multidimensional IRT, which requires more complex multidimensional graphical representations.

Evaluating the Psychometric Adequacy of the Overall Scale as a Unidimensional Scale

Earlier in this chapter, the importance of concurrent modeling of specific and general factors of academic challenges using measurement models such as higher-order and bifactor models was discussed. Nevertheless, the feasibility of deriving an overall score for the scale for

practical use also warrants further examination through classical test theory and an item-level analysis, such as IRT.

An initial validation study revealed that the six subscales of academic challenges exhibited moderate to high correlations with one another (see study 1). Such correlations among the subscales suggest that, although they aim to capture distinct dimensions of a broader construct, these dimensions are not entirely independent and instead share common underlying themes. When factors exhibit moderate to high correlations, a substantial portion of the reliable shared variance among the items can often be explained by a unidimensional factor. This general factor may account for the common underlying themes and shared characteristics across the items. However, when computing non-refined scores such as sum or average scores for a subscale, or when only one or a subset of subscales are modeled, the contribution of this general factor to the items of the modeled subscale(s) is disregarded. Conversely, when calculating a total score for the entire scale or modeling the scale as a unidimensional construct, the unique contribution of each specific factor to their respective items is overlooked.

Ignoring the presence of a general factor or the specific contributions of specific factors can lead to methodological issues and potentially impact the psychometric validity of the scale as a set of discrete measures or as a unidimensional measure. A unidimensional analysis, such as the graded response model, can reveal the relative impact of general and specific factors on the measurement characteristics and reliability of the scales and their resulting scores. This insight allows for the proposal of methods to compute the overall and subscale scores for the scale in practical settings and paves the way for identifying measurement models that more accurately reflect the construct's multidimensionality.

The Added Value of Conducting an Item-Level Analysis in Assessment of Measurement Quality

The benefits of conducting a unidimensional IRT analysis to supplement the previously discussed scale-level examinations of multidimensionality, particularly for the practical application of the Academic Challenges Scale, were discussed. This section expands the focus to explore the broader range of possibilities that IRT can contribute to those factor analytic approaches.

Examining Item Performance at Different Levels of the Latent Attribute

IRT expands measurement modeling beyond test characteristics to incorporate assessments of psychometric validity in terms of person-characteristics (Embretson & Reise, 2013). The assumptions about measurement quality differ between IRT and factor analysis (FA). While FA assumes uniform measurement quality for all individuals, IRT proposes that measurement quality can vary depending on individuals' level of the latent attribute under investigation (Embretson & Reise, 2013). In this dissertation, I use the term *latent attribute* to refer to the underlying factor that each academic challenges subscale aims to measure. I chose this terminology to differentiate from more commonly used terms such as *ability* or *trait*, as these do not accurately capture the nature of academic challenges. Academic challenges are not inherent abilities or fixed traits, but rather represent dynamic and context-dependent experiences that students encounter in their academic pursuits.

In IRT, the probability of endorsing a response is modeled as a function of both person parameters (one's score on the latent attribute) and item parameters (e.g., item thresholds and discrimination). For instance, certain items within the motivation challenges scale might provide more precise measurements (i.e., have a lower standard error of measurement) for individuals

with an average level of motivation challenges compared to those with very high or low levels. The item-level parameters can be aggregated and reported for the entire scale, indicating the extent to which a person's level of motivational challenges can be inferred from their response to certain items and the extent to which the scale comprises items that provide high information across a wide range of motivational challenges. Given the diagnostic objectives of the scale, understanding the distribution range of academic challenges that the scale can measure reliably is crucial. If needed, additional items can be included to cover areas that are under-measured by the scale.

Assessing Response Scale Effectiveness

IRT extends beyond evaluating how well items differentiate between individuals with varying levels of the latent attribute (e.g., low, average, high) and determining the attribute levels required for a given response endorsement. It also allows for a thorough assessment of how effectively the response option categories distinguish between individuals (Ostini & Nering, 2006).

For instance, this capability is particularly valuable in evaluating the 5-point Likert scale used for the academic challenges scales. Through IRT analysis, potential issues with the response categories can be identified. If two adjacent categories show excessive overlap, it might indicate that respondents struggle to differentiate between them, suggesting a potential benefit in merging these categories. Conversely, large gaps between response categories could indicate that the scale fails to capture important nuances in respondents' experiences. This might signal a need to introduce additional response options to ensure the scale adequately captures the full range of the latent attribute. By examining these aspects, we can assess whether the current 5-point scale

optimally captures the range of academic challenges experienced by students, or if modifications to the response scale could enhance its measurement precision and informativeness.

Developing a Shorter Version of the Scales

Valuable insights into the feasibility of developing a shorter version of the scales can be offered by IRT (Embretson & Reise, 2013). To elaborate, through IRT, items that provide high information across a wide range of the latent attribute can be prioritized. A refinement of this nature can be implemented to streamline the scales without significantly compromising their psychometric validity and reliability. The resulting brief scales can be designed to effectively gather essential information, despite their reduced length. This characteristic renders them particularly suitable for practical applications as well as studies requiring frequent data collection, such as those employing experience sampling methods, where the minimization of participant burden is of importance. Consequently, a condensed version of the scales can be developed, which maintains the psychometric integrity of the original scales while reducing the time and effort required for completion by participants.

Evaluating Differential Item Functioning

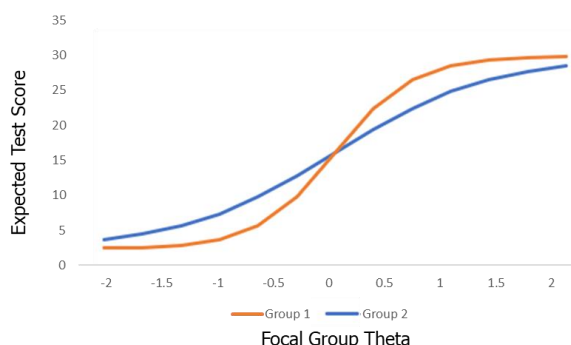
Through the application of IRT, Differential Item Functioning (DIF) can be identified and analyzed. DIF occurs when individuals from different demographic groups (e.g., first-year versus second-year students, or domestic versus international students) exhibit varying probabilities of providing specific responses to an item, despite possessing equivalent levels of the underlying attribute that the scale is designed to measure (Osterlind & Everson, 2009).

The detection of DIF is crucial in ensuring the psychometric validity of the academic challenges scale across diverse student populations. By employing IRT-based DIF analysis, items that may be biased or function differently for certain groups can be identified. This process

allows for understanding how the individual items performs across different demographic categories, potentially revealing hidden biases or measurement inconsistencies that might not be apparent through other analytical methods. For instance, Figure 16 demonstrates a type of group difference that DIF testing in IRT can identify more effectively than invariance testing in factor analysis. As shown, differences between two groups can be nonuniform, meaning that one group may exhibit lower expected scores compared to the other group in specific regions of the latent attribute, while displaying higher expected scores in other regions. When considering a single item, this variability in expected scores across the latent attribute continuum might offset each other when averaged across all respondents within the group. Consequently, indices capable of detecting these types of differences are employed in IRT (Meade, 2010).

Figure 16

An Example of Item Expected Score Plot for An Item with A Non-Uniform DIF



Sample/Parameter Invariance in IRT and Factor Analysis

Invariance is a crucial property in both Item Response Theory and Factor Analysis, but the extent to which it holds differs between the two methodologies. In IRT, invariance refers to the stability of item and person parameter estimates across different samples from the same population (Embretson & Reise, 2000). Parameter invariance refers to the stability of item parameters, such as difficulty and discrimination, across different samples of respondents. When

parameter invariance holds, the item parameters remain consistent, even when the scale is administered to different groups of students. This is important because it ensures that the items function similarly across various subpopulations, allowing for an unbiased measurement.

On the other hand, sample invariance focuses on the stability of ability or attribute estimates (i.e., the estimated levels of academic challenges) across different samples, given the same set of item parameters. When sample invariance holds, students with the same true level of academic challenges will receive similar estimates, regardless of the specific sample to which they belong.

When certain conditions are met, such as having a representative sample, random missingness, ensuring adequate coverage of the range of the latent attribute by the items, and having sufficient information about differential item functioning across demographic characteristics (Rupp & Zumbo, 2006), this invariance ensures that the measurement parameters are unbiased and independent of sample characteristics, allowing for valid application across new samples from the same population. These related properties of IRT allow for the generalization of cut-off scores and the consistent classification of individuals based on their latent trait levels (Hambleton & Swaminathan, 2013). In contrast, factor analysis parameters, such as factor loadings and factor correlations, are often sample-dependent (Brown, 2015). This means that the estimated parameters can vary across different samples from the same population, even when the underlying factor structure is consistent. The sample dependence of FA parameters can be attributed to various factors, such as sampling error, differences in sample size, and violations of assumptions (e.g., normality) (Kline, 2015).

The lower stability of FA parameters across samples compared to IRT parameters is due to fundamental differences in their underlying models. FA uses linear models with scale-

dependent parameters, making them more sensitive to sample characteristics and the variance of items and factors. In contrast, IRT employs non-linear models with standardized scales, designed to produce more stable item and person parameters across different samples (Reise et al., 1993). Additionally, FA does not provide a direct way to estimate the precision of factor scores at different levels of the latent constructs, unlike IRT, which provides information functions to assess measurement precision at different trait levels (Embretson & Reise, 2000; Hambleton & Swaminathan, 2013).

In the context of developing a diagnostic tool such as an academic challenges scale, the sample invariance property of cut-off scores derived from IRT is important. To effectively use the academic challenges scale, it is essential to establish clear cut-off scores that define what levels of academic challenges are considered low, medium, and high. When cut-off scores are sample invariant, it means that the classification of individuals is based on their true level of academic challenges, rather than being influenced by the specific characteristics of the sample used to determine the cut-off scores (Hambleton & Swaminathan, 2013). This invariance allows for the generalization of the diagnostic results across different groups of students, as long as they come from the same population, and enables the comparison of results across different studies or educational settings (Cook & Beckman, 2006), facilitating the accumulation of evidence to support the validity of the interpretations derived from the *Academic Challenges* scale.

In summary, while sample invariance is a desirable property in both IRT and FA, it is more readily achieved in IRT due to the inherent properties of the model. Cut-off scores can be derived in a sample invariant manner in IRT, but the same is not true for FA due to the sample-dependent nature of FA parameters and the lack of built-in mechanisms for deriving cut-off scores linked to specific levels of the latent constructs.

Computerize Adaptive Testing

IRT is a fundamental requirement for developing computerized adaptive scales (Aybek & Demirtasli, 2017). These scales dynamically adjust the items presented to each participant based on their previous responses, rather than presenting the same set of items to everyone. This adjustment is guided by an underlying algorithm that is both theoretically informed and empirically fine-tuned. IRT parameter invariance ensures that the item parameters used by the adaptive algorithm remain stable across different groups of students, allowing for accurate and consistent item selection. Sample invariance, in turn, guarantees that the personalized profiles generated by the adaptive scale are comparable across different samples, as long as the students come from the same population.

The creation of a computerized adaptive scale for measuring academic challenges, particularly with the goal of generating self-diagnostic personalized profiles, is a promising prospect. One area where various SRL measurements differ is the level of granularity at which SRL should be measured (Winne, 2010b). Adaptive scales provide the opportunity to make an individualized balance between the scope (range) and depth (granularity) of measurement. This is achieved by basing the measurement procedure on respondents' answers to a set of screening items and adaptively continuing the process until the model generates a sufficiently broad and detailed profile of the individual's academic challenges. To provide a simplified example, if a student reports relatively high levels of motivational challenges, the algorithm could present items that explore the specifics of these challenges in greater depth. This approach would be more effective than having the student respond to detailed items about an area where they declared they do not perceive much challenge.

Visual Representation of Measurement Quality

A notable advantage of employing IRT, especially in the context of unidimensional IRT models, is its ability to provide graphical representations of measurement quality. These visual depictions offer an accessible means of evaluating and communicating the quality of measurement. Through these graphical representations, the performance of individual items and the scale as a whole can be visually assessed across different levels of the latent attribute. This visual approach allows for a more intuitive interpretation of complex psychometric properties, such as item discrimination and information curves. The insights gained from these visual representations can serve as a foundation for improving the overall quality of the scales. Underperforming items can be easily identified, leading to informed decisions about potential item revisions or, when necessary, the removal of problematic items. This process of visual analysis and subsequent refinement contributes to the development of more robust and psychometrically sound measurement instruments (see section called *Graphs Depicting IRT Parameters*).

Triangulating How Missingness is Handled

The distinct approaches to handling missing data in factor analysis and IRT exemplify the value of cross-validation. IRT effectively bypasses the issue by excluding missing responses from the likelihood function. In contrast, CFA necessitates the use of methods like Full Information Maximum Likelihood (FIML) or multiple imputation to address missing data (Finch, 2008). The alignment or misalignment of the results obtained from these two methods serves to test the robustness of findings in the presence of incomplete data. This difference in addressing missing data illustrates how triangulating the validation study using different

analytical methods can yield a more comprehensive understanding of the scales' measurement quality (Enders, 2010).

In summary, IRT can complement scale-level factor analysis by providing a more granular examination of individual items and assessing measurement precision across different levels of the latent construct. IRT can triangulate the findings from factor analysis, offering a more comprehensive understanding of the scale's psychometric properties. Additionally, IRT can visualize measurement quality in an understandable form, making it easier for researchers and practitioners to interpret and apply the results. By employing IRT, researchers can evaluate the effectiveness of response options, pinpoint items that may be problematic or exhibit differential functioning across various subgroups and develop a more streamlined version of the scale. Moreover, IRT provides sample-invariant cut-off scores that are generalizable and serves as a foundational step in the creation of computerized adaptive scales designed to measure academic challenges, which can offer a more efficient and personalized assessment experience for students.

From Variables to Individuals: Complementary Person-Centered Validation of Academic Challenges Scale

Person-centred versus Variable Centered Approaches

To further investigate the multidimensional structure of the academic challenges scale, a person-centered modeling approach can be used as a complementary analytical technique. Unlike variable-focused methods, person-centered approaches provide a distinct viewpoint by examining the relationships among individuals themselves. Variable-centered modeling focuses on understanding relationships between variables across a sample of individuals. This approach primarily seeks to identify patterns of relationships among variables. It assumes these

relationships are relatively consistent across the sample, allowing sample characteristics to be summarized by average estimations drawn from it (Howard & Hoffman, 2018). Unlike variable-centered approaches that assume a homogeneous population represented by average sample parameters, person-centered approaches view the population as a mixture of distinct latent clusters, requiring cluster-specific estimates to represent the data structure (Howard & Hoffman, 2018). Person-centered analyses like latent profile analysis aim to identify clusters of individuals with similar characteristics that differentiate them from other clusters (Collins & Lanza, 2009; Magidson et al., 2020). These clusters are called *latent* classes or profiles because they are not directly observed but estimated from indicators after accounting for measurement error.

Nevertheless, person-centered modeling is not typically used to investigate psychometric adequacy of constructs. This may be because it is seen as a competing epistemological approach to variable-centered methods. Assessing psychometric quality usually involves exploring psychometric properties of the constructs (variables) being studied, rather than identifying distinct population clusters based on the measure.

While Latent Profile Analysis (LPA) and Common Factor Analysis (e.g., CFA) originate from different epistemological perspectives and address different research questions, their underlying computations share many similarities. In terms of modeling covariance structure, an LPA that groups individuals into K profiles based on their response patterns across a set of observations is mathematically equivalent to a common factor analysis that explains the covariance among the observations using $K-1$ underlying latent factors (Bauer & Curran 2004, Kankara, et al., 2011). LPA and factor analysis both partition covariance among observations into underlying dimensions but do so differently with distinct assumptions. Both approaches model indicators as continuous observations. However, LPA assumes latent dimensions are

categorical (distinct profiles), while CFA assumes dimensions (factors) are continuous (Magidson et al., 2020). That is why despite mathematical equivalence in covariance structure implications, they may differ in interpretability or suitability for specific research questions or data.

Benefits of Conducting Complementary Person-Centered Validation of Academic Challenges Scale

As academic challenges scale is intended for diagnostic purposes, exploring its capability in identifying meaningful clusters of learners with distinct configurations of challenges is beneficial. This exploration can help identify specific groups of struggling students who could benefit from tailored support strategies. Research should determine which levels of challenge or patterns within these clusters are critical and necessitate intervention. Understanding at what point and under what conditions support could be provided is essential for the effective use of the scale in educational settings.

To assess the scale's ability to identify students who may need additional support or benefit most from screening, mean differences in criterion constructs across the identified clusters can be examined. This analysis involves comparing average scores on relevant outcome variables (such as academic performance, engagement, or well-being) across the different clusters identified by the scale. This examination can reveal which specific configurations of academic challenges are most strongly associated with the regulatory processes, engagement, and outcomes, adding to our understanding of the practical utility of how the scale distinguishes between groups of students with varying types and levels of challenges.

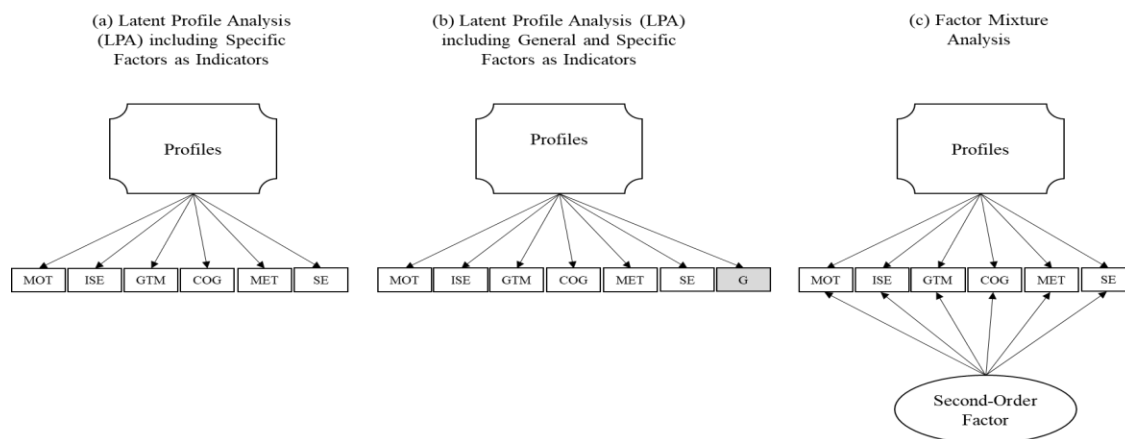
Exploring Academic Challenges' Multidimensionality Through Person-Centered Analysis

Understanding how the population can be decomposed into distinct subpopulations regarding a construct's subscales is valuable for exploring its multidimensionality, especially when it includes an overarching general factor (Morin et al., 2017). Profiles can differ mainly in indicator configurations or represent similar configurations but differing quantitatively in level (e.g., low, medium, high). These two types of profiles are called shape-differentiated and level-differentiated profiles, respectively (Morin et al., 2017).

If latent profile analysis identifies profiles that are primarily differentiated by level differences, the sample can effectively be treated as homogeneous and adequately modeled using variable-centered approaches. If meaningful shape differentiation exists, an individual's score on a variable (e.g., motivation challenge) should be interpreted within the context of their profile membership. Strong shape differentiation across factors may suggest more complex factor interactions than can be captured by a covariation structure represented by a general factor.

Analyzing shifts in profile level and shape when incorporating the general factor provides valuable insights into the multidimensional nature of the scale. This aim can be fulfilled by comparing latent profiles based on specific factors alone with latent profiles that include both general and specific factors. The objective is to test how the addition of the general factor impacts the scale's ability to partition the sample into meaningful clusters of participants. The general factor, representing common item variance, which captures common item variance, enabling specific factors to further contribute to shape differentiation.

Two latent profile analyses, illustrated in Figures 17a and 17b, can be conducted to examine the multidimensionality of the academic challenges scale.

Figure 17*Person-Centered Mixture Models for the Academic Challenges Scale*

Note. Abbreviations: MOT: Motivation, ISE: Initiating & Sustaining Engagement; GTM: Goal & Time Management; COG: Cognitive, MET: Metacognitive Challenges; SE: Social & Emotional Challenges; G: General.

The first LPA (model a) uses factor scores from the best-fitting first-order model (e.g., First-order ESEM) as profile indicators. The second (model b) employs factor scores from the best-fitting bifactor model (e.g., Bifactor ESEM) as profile indicators. As previously noted, in higher-order constructs, first-order and higher-order constructs are not properly separated and share overlapping variance (Chen et al., 2006). Consequently, this comparison is only viable between a first-order and a bifactor solution.

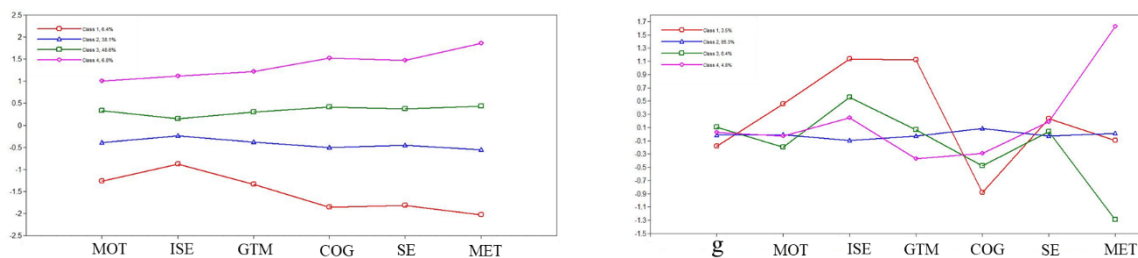
Another potential person-centered alternative to LPA is Factor Mixture Modeling (Figure 17c), which is not appropriate for the objective discussed for several reasons. Firstly, this model should incorporate factor scores from the best-fitting first-order factor model, with a latent general factor added to account for commonalities among profile indicators. However, this general factor only captures the residual covariance unexplained by latent profiles, not all covariance among indicators. More crucially, for model identification purposes, the mean of this general factor must be constrained to equality across profiles. This constraint limits its utility in examining the impact of using coexisting general and specific factors as profile indicators. In

other words, while it allows for partial separation of general and specific factors, it assumes all profiles are invariant in terms of the general factor's level (Morin et al., 2017; Morin & Marsh, 2015).

Comparing first-order and bifactor LPA models (models a and b) offers insight into how the inclusion of the general factor might absorb level differences across profiles, potentially enabling additional qualitative differentiation of profiles in the second model (model b) compared to the first (model a). Figure 18 illustrates a scenario where including the general factor yielded profiles in which the general factor explained a considerable portion of the quantitative differentiation between profiles (level), while specific factors accounted for more shape-related differentiation (Morin et al., 2017).

Figure 18

An example of Extracted Profiles Based on First-order and Bifactor Indicators



Note. The left plot represents the extracted profiles when the factor scores from the first-order model were used as the indicators of the profiles. The right plot represents the extracted profiles when the factor scores from the bifactor model were used as the indicators of the profiles. As demonstrated, the first-order profiles primarily differ in terms of level. However, after the addition of the general factor (g), the main differences between profiles shift to differences in shape.

In sum, the primary purpose of conducting this clustering analysis is to explore structural complexities that might challenge the appropriateness of traditional variable-centered analyses for modeling the construct's multidimensionality. Using person-centered analyses as a complementary method can reveal the degree of oversimplification that forcing a one-size-fits-all variable-centered model may have imposed on a potentially heterogeneous sample. Specifically,

the existence of complex shape differentiation among first-order factors, if observed, would challenge the validity of interpretations derived from the general factor as a representation of covariation among observations and factors. Such a scenario would suggest the need for more complex or even non-linear structures to adequately model what the construct's dimensions represent in combination. However, the level differentiation among first-order factors may indicate that relationships between indicators could be effectively modeled using variable-centered analyses, without necessarily resorting to mixture modeling such as latent profile analysis.

On the other hand, this complementary analysis highlights which indicators should be employed in clustering studies. For example, it can reveal the biases in clustering that omitting the general factor as an indicator can introduce in clustering studies using the academic challenges scale. Put differently, while variable-centered methods can guide how clustering analysis should be conducted, person-centered analyses also inform the multidimensional structure of the scale based on the degree of heterogeneity in the configurations of profile indicators within the population. This, in conjunction with an examination of how distinct clusters are related to criterion outcomes, carries significant implications for a measure with diagnostic objectives, such as the Academic Challenges Scale.

Summary of the Chapter

In this chapter, the significance of examining the scale-level architecture of academic challenges scale was initially addressed. It was also noted that this top-down approach can be complemented by a bottom-up, item-level analysis of measurement precision, particularly when validating the scale for practical applications. Furthermore, considering the diagnostic aspects of the scale and the hypothesis that the multidimensional structure of the Academic Challenges

Scale necessitates a hierarchical structure, the benefits of validating it using person-centered clustering methods were discussed. The combination of these approaches offers a comprehensive understanding of the measurement quality of the Academic Challenges Scale for both research and practical purposes. In the upcoming chapter, a tabular overview will be presented, summarizing the objectives and contributions of each analytical method in illuminating different aspects of the measurement quality of the scale.

Chapter 3: Research Purpose, Context, and Manuscript Overview

“Data is like garbage. You’d better know what you are going to do with it before you collect it” attributed to Mark Twain

Research Purpose

The purpose of this dissertation was to investigate the psychometric properties of the *Academic Challenges* scale. This examination was carried out through three interconnected studies, each addressing specific aspects of the scale's psychometric properties. Table 6 provides a detailed overview of the main purpose and research questions for each study.

Study 1, a collaborative study (Hadwin, Rostampour, Winne, submitted for publication), introduced and examined four psychometrically separate scales that collectively form the Self-Regulated Learning Profile and Self-Diagnostic tool (SRL-PSD). These scales measure:

(1) students' self-efficacy for SRL, (2) perceived importance of SRL, (3) SRL practices, and (4) perceived academic challenges experienced during studying.

Studies 2 and 3 expanded on Study 1 by focusing on specific aspects of the *Academic Challenges* scale, examining it from two perspectives. Study 2 concentrated on item-level precision, employing Item Response Theory (IRT) to examine how well individual items measure the intended constructs across various levels of academic challenges. Study 3 explored the multidimensionality of the *Academic Challenges* scale, investigating the potential existence of an overarching factor of academic challenges and the impact of incorporating cross-loadings within the specified measurement model.

In other words, these latter two studies extended the initial study by providing a more nuanced understanding of the *Academic Challenges* scale's psychometric properties. I focused exclusively on this scale to maintain a manageable scope for the current dissertation, while acknowledging the necessity of examining the other three scales in future research projects.

Table 6*Main Objectives and Research Questions of the Dissertations Studies*

Objectives	Analytical Methods	Samples
Study 1 Validating a newly developed SRL Profile and Self-Diagnostic instrument (SRL-PSD) by examining its psychometric properties and criterion utility among post-secondary university students. RQ1. Are the four scales of the SRL Profile and Self-Diagnostic instrument (SRL-PSD) psychometrically adequate? RQ2. Do the four scales of the SRL-PSD have predictive utility for understanding academic outcomes? RQ3. How are the academic challenges and SRL practices scales of the SRL-PSD related?	• EFA, CFA for psychometric examination • SEM for Structural Models testing the criterion utility	• Pilot sample= 220 undergraduates enrolled in a learning to learn course • Main Sample:473 undergraduates across various disciplines
Study 2 Examining the psychometric properties of the items that constitute the <i>Academic Challenges</i> scale across various levels of the intended constructs. RQ1. How precisely do the items within each specific scale of the <i>Academic Challenges</i> scale measure different levels of their target challenges as latent unidimensional constructs? RQ2. How precisely do all items combined in a single scale measure different levels of a latent general academic challenges construct? RQ3. To what extent do the items in the <i>Academic Challenges</i> scale exhibit differential item functioning across the five participating university sites? RQ4. How do average and factor scores for specific scales, and for the overall scale align with their corresponding true factors? RQ5. How are specific scales, the overall scale, and composite scores of the academic challenges scale associated with students' GPA expectations, GPA Self-Efficacy, and end-of-semester GPA?	• Samejima's Polytomous Graded Response IRT • Mokken scaling for investigating monotonicity	2,671 university students from five Canadian universities across 4 provinces (cohort 1: 1245 cohort 2: 1426)
Study 3 Addressing potential limitations in previous measurement models of the <i>Academic Challenges</i> scale, specifically by exploring the existence of a general factor of academic challenges and allowing for cross-loadings in the specified measurement model. RQ1. Which measurement model best represents the multidimensionality of the <i>Academic Challenges</i> scale? RQ2. How does modeling an overarching academic challenge factor and allowing for cross-loadings impact the interpretation of the <i>Academic Challenges</i> scale? RQ3. What are the reliability and dimensionality characteristics of coexisting general and specific factors of academic challenges? RQ4. How do the different measurement models relate to criterion variables (GPA expectations, GPA confidence, and end-of-semester GPA), and what does this reveal about their predictive utility? RQ5. To what extent does the best-fitting measurement model demonstrate psychometric invariance across year of study, sex at birth, and cohort intake?	• A sequential strategy including Correlated, Second-order &, Bifactor CFA, and Correlated, Second-order, & Bifactor ESEM. • Ancillary Measures of Reliability for the Bifactor Model	

Research Context and Ethics

Ethics Statement for Study 1

This study received ethics approval as part of broader research initiatives at the University of Victoria focusing on student success, self-regulated learning, and promoting adaptive regulation. The overarching project was titled "Promoting Adaptive Regulation through Innovative Technologies".

Participants were recruited through three main channels, (a) the Psychology department's research participation pool and (b) an elective educational psychology learning-to-learn course and (c) an online strategy library website only accessible to the enrolled students.

All participants were enrolled students who provided informed consent. They were fully briefed on the study requirements and informed of their right to withdraw from the research at any point during the semester. In the event of a withdrawal request, all identifiable data, including performance indicators collected up to that point, were destroyed and excluded from the research.

Ethics Statement for Study 2 and Study 3

Ethics approval was obtained for these studies as a part of a larger research projects aimed at delivering specialized workshops designed to help students learn and develop useful skills, set long-term goals and channel their unique personalities toward achieving those goals (<https://www.univentureproject.org/>). This research project is funded through Social Sciences and Humanities Research Council (SSHRC) Partnership Grant [Grant#: 895-2019-1021]. The project is ethically approved by the five participating data collection sites by their respective ethic review boards [Dalhousie University (2021-5689), Saint Francis Xavier University (ROMEO File #25486), University of Montreal (2020-883), York University (2021-272) and University of British Columbia – Okanagan (H21-02007)]. Students' consent to access their institutional data from the registrar's office of their universities was obtained separately from their general consent to participate in the study. All non-consenting students were removed from

the dataset. The dataset was fully anonymized, cleaned, and shared across sites after evaluating all ethical considerations and the existing regulations enforced by the relevant Research Ethics Boards.

Pandemic Context

Data for the studies were collected in Fall 2020, Winter 2021, Fall 2021, and Winter 2022 during a unique and unprecedented pandemic. During this period, all participating students engaged in university courses via online, off-campus education. For the majority of both students and instructors, online instruction was a relatively new mode of learning and teaching. The students were geographically dispersed across Canada and internationally, with most studying from their homes. The extent of COVID-19 restrictions on socializing varied both within Canada and around the world. Furthermore, many extracurricular activities were either cancelled or faced substantial limitations. These unique circumstances should be taken into consideration when interpreting the study's results.

Participants

Four student samples were used in the studies included in this dissertation. The same dataset was used for both Study 2 and Study 3. This decision was made because these two studies are complementary in nature: one focuses on the precision of items, and the other on the multidimensional structure of the measure. These distinct focuses provided a complementary understanding of the scale by maximizing insights through different analytical lenses.

Additionally, full measurement invariance was observed between the two cohorts of the sample (Fall intake $N=1245$, Winter intake $N=1426$). Findings from Study 3 supported *configural*, *metric*, *scalar*, *strict*, latent variance-covariance, and latent means invariance across these two cohorts. This indicated that even if the analyses were conducted with different cohorts (e.g., one cohort for Study 2 and another for Study 3), the findings would have been consistent.

Using the same sample for the two studies allowed for examining the methodological synergies of the analytical methods employed. Furthermore, Study 1, which used two separate samples, yielded comparable results. This consistency across different samples lends credence to the reliability of the findings and suggests that the results may be applicable to analogous populations.

Participant Information for Study 1

Study 1 Pilot Sample

A pilot sample of 220 university students enrolled at a Western Canadian university participated in the initial development and exploratory factor analysis (EFA) of the scales in Study 1. These students were enrolled in an introductory Learning to Learn course and represented a diverse range of programs and academic levels. The data collection occurred during week 10 of a 13-week academic semester, with students completing a battery of questionnaires for self-assessment as part of their course requirements. The sample demographics were as follows:

- Sex at Birth: 48% female
- Age: Mean: 20.1 years (SD = 2.5, range = 18-49)
- Mean GPA: 5.2 out of 9 (equivalent to a B grade; SD = 1.9, range = 1-9)
- Year of Study: 76% first-year, 14% in other years
- Enrollment Status: 89% full-time, 11% part-time
- Residency: 86% Canadian citizens, 3% permanent residents, 11% international students
- The students represented 8 faculties, including education, social science, engineering, computer sciences, fine arts, humanities, and business.

The learning course was a credit-bearing, elective offering primarily designed for first-year undergraduate students from various faculties and achievement levels across the university. This course was not a remedial program for struggling students but rather a course for a wide

range of students at all achievement levels interested in strategies for university success. It covered a comprehensive range of topics crucial for academic success, including self-regulated learning based on Winne and Hadwin's (1998) SRL model, motivation, learning and memory processes, effective reading techniques, diverse learning strategies, notetaking skills, and exam preparation methods. It is important to note that the pilot sample was used exclusively for the initial piloting of the scales and conducting EFA. Main analyses in Study 1 were performed using a separate main sample.

Study 1 Main Sample

The main sample consisted of 473 undergraduate students enrolled at a Western Canadian university. Participants were recruited through three channels during week 11 of a 13-week academic semester in the Fall of 2020. Data collection was conducted via LimeSurvey (Version 3.2, 2020). The recruitment channels were the psychology department's research participation pool ($n = 354$), an elective educational psychology learning-to-learn course ($n = 88$), and a digital learning strategy library available university-wide ($n = 31$). The sample demographics were as follows:

- Sex at birth: 74% female
- Age: Mean 20.4 years ($SD = 3.3$)
- Academic Year: 39% first-year students, 61% in other years
- Faculty Distribution:
 - Social Sciences (including Psychology): 55%
 - Science: 17%
 - Engineering: 9%
 - Other faculties (Humanities, Business, Education, Fine Arts, etc.): 19%
- Citizenship Status:
 - Canadian Citizens: 89%
 - International students: 8%
 - Permanent Residents: 3%
- Mean GPA: 6.6 out of 9 (equivalent to B+ to A-; $SD = 1.8$)

Incentives varied based on the recruitment channel. Students from the learning-to-learn course received no additional incentives, as the data collection was integrated into regular course activities. Strategy library participants were entered into a draw for an Amazon gift card. Participants from the psychology department's research pool were awarded 1.5 course credits for completing the survey. These students were also invited to attend a 15-20 minute debriefing session via ZOOM at the semester's end.

Participant Information for Studies 2 and 3

The data for the study were collected as part of the UniVenture program (<https://www.univentureproject.org/>), a wellness program tailored for university students. The program focused on offering specialized workshops that equipped students with valuable skills, facilitate the development of long-term goals, and guide them in leveraging their unique personalities to effectively pursue their goals.

Studies 2 and 3 used baseline data collected during weeks one to five of the Fall 2021 and Winter 2022 semesters via REDCap electronic data capture tools (Harris et al., 2019). The data comprise university students from five Canadian universities across four provinces (Nova Scotia, Ontario, Quebec, and British Columbia). Institutional data, such as GPA, was obtained from the universities' registrar's offices. The sample consisted of N=2671 participants (Fall intake=1245, Winter intake= 1426), with 65% being female, 32% male, and 3% non-binary. Among the participants, 56% were first-year students, while 44% were in their second year of study. Students' age ranged from 18 to 25 (Mean=19, SD=1.5).

The sample represents a diverse range of academic programs, with the largest groups being Science (32.2%), Arts (19.9%), Business/Commerce (8.6%), Physical Education/Kinesiology (8.5%), and Engineering Applied Science (6.5%). Together, these five categories account for about 75.7% of the valid responses. The remaining 24.3% is distributed

across various other programs including Nursing, Computing, Education, Music, and Fine Arts, as well as an "Other" category that comprises 15.8% of responses.

In terms of diversity and student status, 27.2% of participants identified themselves as visible minorities, while 4.1% preferred not to answer this question. The sample included 14.8% international students. Regarding self-reported employment status, 11.6% of the students worked full-time, 37.8% worked part-time, and 48.3% did not work alongside their studies.

During the baseline phase of data collection, which served as the data source for the current study, students provided responses to a range of topics including demographics, personality, substance and alcohol use, anxiety, and depression prior to the academic challenges items. As an incentive for completing the study, participants could choose to receive either a \$20 Amazon gift card or credit towards research participation.

All analyses were subjected to measurement invariance testing to guarantee uniformity across the participating universities, as well as to ensure consistency between the two cohorts under study: Fall 2021 and Winter 2022 intakes. After the data collection at baseline, about ten percent of students attended the program workshops. Given the potential influence of workshop attendance on the GPAs collected at the end of academic semesters, a sensitivity analysis was performed to confirm there was no statistically detectable difference in GPA between students who attended the workshops and those who did not.

A retrospective power analysis indicated that a model with $\alpha = 0.05$, $df = 269$, and $N = 2671$ would yield acceptable power ($1 - \beta = 0.99$), being satisfied by the current sample size (Jobst et al., 2023).

Measures

The SRL-PSD Instrument (Study 1)

The SRL-PSD scales evolved through iterative refinement over several years of data collection and pilot testing with post-secondary students. Additional details are provided for each scale in subsequent sections. The complete set of items is available in the Appendix. Participants were instructed to base their responses on their studying experiences from the past two weeks. This approach aimed to capture studying beliefs, practices and challenges across diverse tasks and subjects while minimizing forgetting and overgeneralization that could occur with longer reference periods.

Self-Regulated Learning Practices (*SRL-P*)

The SRL-Practices Scale (SRL-P) assesses students' self-reported engagement in actions aligned with SRL processes. It consists of 31 items across 8 subscales: (1) *task understanding*, (2) *goal management*, (3) *appraisal of task value*, (4) *motivation appraisal*, (5) *time management*, (6) *metacognitive monitoring*, (7) *metacognitive adaptation*, and (8) *academic social engagement*. Participants rated their agreement with each practice over the past two weeks using a five-point Likert scale from (1) strongly disagree to (5) strongly agree, with higher scores indicating more proactive SRL behaviors.

Items selection was guided by Winne and Hadwin's (1998) model of studying as self-regulated learning, prioritizing (a) metacognitive practices often underrepresented in previous measures, (b) strategic actions during studying rather than task-specific techniques, and (c) self-regulatory processes across SRL phases. Items were drawn from various sources, including an earlier regulated learning scale (McCardle & Hadwin, 2015), regulation of motivation items (Kim et al., 2020), time management and procrastination items (Won & Yu, 2018), and social

belongingness and engagement practices (Won et al., 2021). The latter were included to address social and emotion-supporting practices, particularly relevant during pandemic-driven online learning. Additionally, items were compared to weekly open-ended descriptions of strategies students used to address academic challenges that impeded their goal attainment or progress toward weekly self-set study goals.

SRL Importance Scale (SRL-I)

The SRL Importance scale (SRL-I) consists of 10 items describing broad categories of SRL practices, such as planning. These items were selected to represent the most salient aspects of the SRL-practices scale. Participants rated the importance of each item for their studying over the past two weeks using a 5-point Likert scale, ranging from (1) very low priority to (5) very high priority. Higher scores indicated that students considered these SRL practices as important priorities during their study sessions.

SRL Self-Efficacy Scale (SRL-SE)

The SRL self-efficacy scale (SRL-SE) uses the same 10 items from the *SRL Importance* scale. However, in this context, students were asked to rate their confidence in successfully engaging in these SRL practices over their upcoming two weeks of studying. Ratings were made on a 5-point Likert scale, ranging from (1) not confident to (5) completely confident. Higher scores on this scale indicated that students possessed high self-efficacy or confidence in their ability to successfully implement SRL practices during their future study sessions.

Academic Challenges

As this scale is also used in Studies 2 and 3, it will be detailed as one of the measures used in these studies.

Academic Performance (Study 1)

Academic performance was measured using end of semester registry GPA gathered from institutional records. Semester GPA represents performance across all courses in that academic semester and is reported on a 9-point scale where 0 = F(0-48%), 1 = D (50-59%), 2 = C(60-64%), 3 = C+(65-69%), 4 = B-(70-72%), 5 = B(73-76%), 6 = B+(77-79%), 7 = A-(80-84%), 8 = A(85-89%), and 9 = A+(90-100%).

Academic Challenges Scale (Studies 1, 2, & 3)

The Academic Challenges scale is one of four measures included in the Self-Regulated Learning Profiles and Self-Diagnostic tool. Together, these four measures provide a comprehensive and relational profile of students' self-regulated learning.

The Academic Challenges scale, which can be found in the appendix, underwent an iterative development process involving multiple rounds of data collection and pilot testing with students in post-secondary education (e.g., Hadwin et al., 2019). Initially, students were asked to document challenges encountered during their weekly study sessions throughout a semester, aiming to capture direct insights from their experiences. These responses were then analyzed qualitatively and organized into a dropdown list, providing a structured method for subsequent students to identify encountered challenges. An “other” option was included, offering an open text field for unlisted challenges, ensuring the list's comprehensiveness in reflecting students' actual experiences. The list was refined and expanded until a saturation point was reached, characterized by a negligible rate of uncategorized challenges reported over two consecutive semesters.

It was hypothesized that the challenges could be categorized into five principal areas linked to student success (e.g., Bowden et al., 2019; Wong et al., 2024): *motivation, behavior, cognition, metacognition, and social-emotional challenges*. The first version of the scale was

developed based on this conceptual model, and its psychometric properties were tested in Study 1 using two samples of university students.

Findings from Study 1 led to a reconfiguration of the measurement model, resulting in 31 items across six subscales. These subscales include: (1) *motivation*, (2) *initiating and sustaining engagement*, (3) *goal and time management*, (4) *cognitive*, (5) *metacognitive*, and (6) *social and emotional challenges*.

These subscales assess the extent to which students perceive various study-related challenges. Participants reported on their experiences over the previous two weeks, rating their agreement with each challenge on a five-point Likert scale ranging from (1) strongly disagree to (5) strongly agree, with disagree (2), neutral (3), and agree (4) in between. The selection of a two-week timeframe aimed to strike a balance between capturing recent study experiences and ensuring an adequate range of recallable incidents. High scores indicated perceived struggles with aspects of studying that are theoretically and empirically linked to student success and performance. The six academic challenge subscales are described below.

Motivation Challenges (MOT)

The MOT scale is defined by 3 items: struggling with “believing I can do my work”, “feeling like my work was worth doing”, and “persisting when things got tough”. These items encompass three key aspects of students’ motivational challenges: belief in one’s abilities (self-efficacy), perceived value and meaningfulness of the work (value), and capacity to continue efforts in the face of difficulties (persistence). Collectively, these elements reflect the motivational foundations crucial for successful and sustained engagement in educational activities (Eccles & Wigfield, 2020; Pintrich, 2004; Schunk et al., 2013; Wigfield & Eccles, 2000; Winne & Hadwin, 2008).

Initiating and Sustaining Engagement Challenges (ISE)

The ISE scale comprises 4 items that measure the internal motivation to start and persist with tasks, as well as the ability to manage attention. This factor includes difficulties in sparking interest in schoolwork (“wanting to do my schoolwork”), in overcoming procrastination (“putting things off until the last minute”), in handling external and internal distractions, and in maintaining focus and concentration (“managing distractions”; “focusing or concentrating”). ISE captures the challenges related to both the initial drive to engage in academic activities and the sustained attention needed to remain engaged and productive.

The Goal and Time Management Challenges (GTM)

The GTM scale is defined by 4 items that measure struggles with effective time and task management (“Managing my time, tasks, or goals; Organizing time and tasks”), pursuing academic goals (“Keeping my commitments or goals”), and adhering to deadlines (“Meeting deadlines”). GTM reflects a student's challenges in organizing and managing their academic workload efficiently and consistently, which is crucial for successful academic performance and productivity (Wolters & Brady, 2020).

Cognitive Challenges (COG)

The COG scale encompasses 6 items that measure struggles with cognitive processes involved in learning and understanding. This factor includes difficulties with recalling learned material (“remembering things that I learned”), comprehending and internalizing information (“understanding learning materials and concepts”; “identifying important ideas”), analyzing and synthesizing new concepts (“elaborating and connecting what I learned”; “applying my learning”), and communicating effectively (“language and communication”). Overall, the COG

scale reflects struggles with the cognitive demands placed on students in processing, retaining, and applying academic content.

Metacognitive Challenges (MET)

The MET scale includes 8 items measuring challenges related to the ability to plan, monitor, and adjust one's approach to learning. It reflects struggles with the higher-order thinking required to understand tasks and facilitate task engagement (“breaking tasks down,” “interpreting tasks and expectations,” “prioritizing or figuring out what is important”), set appropriate learning goals (“setting goals for studying”), monitor learning (“monitoring my task or learning progress”), and apply suitable strategies to optimize learning (“knowing how to tackle tasks”; “knowing how or when to fix strategies/study skills”; “applying strategies/techniques properly”). Metacognitive regulation, which includes metacognitive knowledge, awareness, monitoring, and control, is fundamental to self-regulated learning (Efklides, 2006, 2008; Panadero, 2017; Winne, 2011, 2017; Winne & Hadwin, 1998).

Social and Emotional Challenges (SE)

The SE scale comprises 6 items measuring struggles with interpersonal and emotional aspects of the student experience. It includes the ability to form connections and maintain relationships (“feeling connected”, “feeling lonely”, “managing relationships”), find joy in their academic environment (“finding enjoyment at university”), and manage emotional well-being (“managing my emotions/feelings”; “taking care of my mental health and well-being”). In line with the relevant literature (e.g., Beaumont et al., 2023; Davis & Hadwin, 2021; Pekrun & Linnenbrink-Garcia, 2012), the SE challenges highlight the significance of emotional well-being and sense of belonging (Won et al., 2021) in overall academic success and personal development. This subscale was added during the pandemic, a period characterized by limited

interpersonal interactions due to restrictions and the widespread adoption of online instruction modes. Awareness of the broad spectrum of challenges, including emotional and social aspects, supports the shift towards holistic educational models that simultaneously foster well-being and learning skills (White, 2023).

Criterion Measures (Studies 2 & 3)

GPA Expectation (GPA-Exp)

GPA expectation, measured by a single item, assessed students' anticipations regarding their academic performance for the semester or term in which they were enrolled. This item asked participants to indicate the overall Grade Point Average (GPA) they realistically expect to achieve. The response options were structured on a scale to capture a range of expectations, from highest to lowest: High GPA, Strong GPA, Satisfactory GPA, Low GPA, and Failing GPA. Each descriptor is accompanied by an institution-specific quantified GPA to clarify what the descriptors mean in terms of actual grades at each institution (e.g., Strong GPA (2.7 - 3.7)).

GPA Self-Efficacy (GPA-SE)

GPA Self-Efficacy, measured by a single item, was designed to gauge students' confidence levels regarding their ability to achieve their expected GPA. After students specified their expected GPA using the provided descriptors and quantified ranges, they were asked to assess their confidence in attaining that GPA. The response options for this item ranged from *not at all confident* to *extremely confident*, with *neutral* serving as the midpoint. This measured students' self-efficacy towards their academic performance for the semester they were enrolled in. Together with GPA expectation, this construct helped capture the psychological aspect of students' academic expectations, adding depth to the understanding of their academic self-perception.

End of Semester GPA (Studies 2 & 3)

End-of-semester GPAs for the students were obtained directly from the registrar offices of each institution. As each university employed its own GPA measurement scale, it was necessary to convert the GPAs to a common scale. This conversion process entailed aligning the unique grading systems of each university to a standardized 5-point letter grade GPA scale. Numerical grades were converted to their equivalent letter grades based on conversion guides supplied by each institution. The choice to employ a 5-point letter grade scale was influenced by its status as the most detailed scale used by two of the participating universities. Standardizing the GPAs on a common scale ensured consistency and facilitated meaningful comparisons across the different universities and their respective grading systems. On this converted scale, an *A* grade signifies “Excellent”, a *B* grade indicates “Very Good”, a *C* grade represents “Satisfactory”, a *D* grade signifies “Barely Passed”, and an *E* grade indicates “Failed”.

Overview of Manuscripts

Manuscript 1: Advancing Self-reports of Self-regulated Learning: Validating New Measures to Assess Students' Beliefs, Practices, and Challenges (Appendix A)

Background: Self-regulated learning (SRL) plays a vital role in academic success. There is a need for comprehensive, psychometrically sound measures of students' perceptions of SRL that are applicable across various academic contexts.

Objective: To test the psychometric adequacy and predictive utility of four separate scales measuring students' perceptions of SRL: self-efficacy, importance, practices, and challenges experienced during studying.

Methods: The study involved two samples: a pilot sample of 220 and a main sample of 473 post-secondary students from various academic disciplines. Confirmatory factor analyses were conducted on four scales: a 4-factor SRL self-efficacy scale, a 4-factor SRL importance scale, an 8-factor self-regulated learning practices scale, and a 6-factor academic challenges scale. Predictive utility was examined through associations with academic performance (end of semester GPA).

Results: Confirmatory factor analyses supported the psychometric adequacy of all four scales. Predictive utility analyses revealed that (a) 5 out of 6 academic challenges were inversely associated with academic performance, (b) motivation challenges had incremental power in predicting academic performance above and beyond other challenges, and (c) each academic challenge was associated with select SRL practices, with task understanding and time management practices showing the highest associations with challenges.

Conclusion: This study validated a comprehensive SRL instrument that addresses the practical need for measures applicable across various educational contexts. The instrument can be used by educators to assess SRL interventions, by researchers to contextualize behavioral traces of SRL, and by learners to identify areas for SRL improvement.

Manuscript 2: Academic Challenges under the Lens of Item Response Theory: An Item Level Analysis of Measurement Precision in Perceived Academic Challenges Scale (Appendix B)

Background: Students face various academic challenges that may impact their learning in potential positive or negative ways. Previous research has validated a self-report measure of these perceived challenges using confirmatory factor analysis. Given the diagnostic objective of the scale, there is a need for more precise item-level examination across various ranges of the intended constructs.

Objective: To examine the psychometric properties of the *Academic Challenges* scale using a polytomous Graded Response Model Item Response Theory (IRT), and to assess the adequacy of the measure's arithmetically computed scores for practical application.

Methods: The study included 2671 university students from five North American universities. IRT analyses were conducted on six dimensions of perceived academic challenges (motivation, initiating and sustaining engagement, goal and time management, cognitive, metacognitive, and social-emotional challenges) and on an overall scale. The psychometric properties of composite scores were evaluated, and the measure's relationship with three criterion variables—GPA expectations, self-efficacy for GPA, and actual end-of-semester GPA—was investigated.

Results: IRT analyses confirmed the precision of most items in assessing various levels of intended constructs and identified imprecise items for revision. The measure demonstrated adequate psychometric properties, including appropriateness of its total score, and showed relevance to criterion constructs, supporting its potential practical utility.

Conclusion: The study supports the psychometric adequacy of the *Academic Challenges* scale. Emphasizing the importance of continuous refinement of self-report measures in educational research, recommendations for improving future iterations of the scale were provided.

Manuscript 3: Exploring the Multidimensionality of the Perceived Academic Challenges Scale: A Bifactor ESEM Approach (Appendix C)

Background: Academic challenges are highly interwoven. Previous psychometric research on the *Perceived Academic Challenges* scale has not examined the potential existence of an overarching factor representing overall academic challenges, nor has it considered the potential influences that items may receive from factors other than their designated *a priori* factors. These limitations could have possibly reduced the ecological validity of the measure.

Objective: To investigate the multidimensionality of the *Perceived Academic Challenges* scale using a sequential strategy that allows for examining the potential existence of a general factor and incorporating cross-loadings within a general factor, both individually and in combination.

Methods: The study involved 2,671 university students from five Canadian universities. Bifactor and second-order measurement models were used to investigate the existence of an overarching factor. Exploratory Structural Equation Modeling (ESEM) was employed to incorporate cross-loadings. The scale's predictive utility was examined through associations with students' GPA expectations, self-efficacy for GPA, and actual end-of-semester GPA.

Results: Findings supported the presence of a general factor underlying the responses and the appropriateness of incorporating cross-loadings to improve the overall fit of the measurement model. A bifactor ESEM approach, which incorporated both a general factor and cross-loadings, demonstrated the best fit to the data and showed the strongest associations with criterion variables. Given the substantial overlap between the general factor and the metacognitive challenges scale, modifications to the bifactor ESEM model were made to anchor the general factor to the metacognitive challenges using a S-1 bifactor ESEM model.

Conclusion: Modeling the *Perceived Academic Challenges* scale benefits from advanced measurement models capable of improving its ecological validity. The bifactor ESEM approach provided a more nuanced understanding of the scale's structure and its relationship with academic outcomes. Suggestions for further improvements to the scale are offered, emphasizing the necessity for continual refinement of measurement tools in educational research.

Chapter 4: Discussion

“Not everything that can be counted counts, and not everything that counts can be counted” - William Bruce Cameron

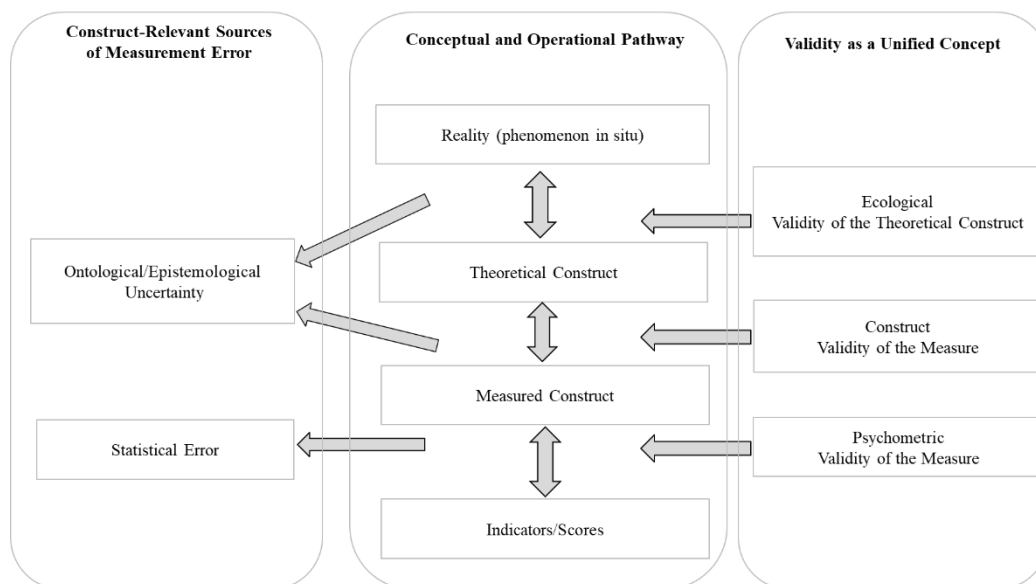
In this dissertation, I aimed to enhance the understanding of the Academic Challenges scale's psychometric properties. This chapter employs a structured approach to discuss the findings. I begin by introducing a conceptual framework that positions the investigation of psychometric properties within the broader context of evaluating qualities of measurements. Following this brief conceptual discussion, I summarize the findings from studies on the scale's multidimensional structure, person-centered clusters, and item-level precision. I then provide suggestions for improving the measure and methods for computing scores for practical applications. Where applicable, I present concise overviews of findings and recommendations in tabular format. The chapter concludes with recommendations for future research directions and addresses limitations of the conducted research.

Measurement Quality Versus Psychometric Adequacy

Figure 19 presents a conceptual framework for understanding the relationship between different subtypes of validity and sources of error in the measurement process.

Figure 19

Measurement: Linking Reality to Scores



It depicts the bidirectional flow from reality to the indicators/scores of the measure, passing through the stages of the theoretical construct and the measured construct.

The framework addresses several key aspects of validity. Ecological Validity (Bronfenbrenner, 1977) of the theoretical construct assesses how well the theoretical construct reflects the real-world phenomenon. Construct Validity (Cronbach & Meehl, 1955; Flake et al., 2017) of the measure examines the alignment between the measured construct and the theoretical construct. In other words, it measures how closely the measured construct approximates the conceptual definition of the construct, while being subject to influences from both ontological/epistemological uncertainties and statistical error. The psychometric validity of a measure refers to the degree to which evidence supports the accuracy of the measure as a representation of its intended indicators/scores.

The framework also considers sources of error. Ontological/epistemological uncertainty highlights the reductionism due to inherent complexity of real-world phenomena and/or to epistemic boundaries of the knower, both impacting the theoretical constructs developed. Statistical error refers to random aspects of indicators that fail to accurately contribute to the representation of the measured construct. Measurement error results from both statistical error and ontological/epistemic uncertainties. However, it is often equated with statistical error, overshadowing the significance of the latter. This might be partly because empirically identifying characteristics of indicators that do not contribute to the total score is more straightforward than theoretically investigating the reductionism imposed by the nature of the known and the limits of the knower in defining the theoretical and measured constructs.

A measure can meet conventional criteria for psychometric validity (e.g., satisfactory fit indices) because a large proportion of item covariances can be attributed to the latent structure.

However, a measure might be judged to have low construct validity if it lacks important facets of the theoretical construct or violates major theoretical assumptions (e.g., temporal/contextual characteristics). Similarly, a measure can be judged to have high levels of construct validity and psychometric adequacy yet fail to validly represent something in the real world if the theoretical model underlying the measure is overly reductionistic.

Therefore, the quality of the measurement should not be solely based on its psychometric properties, as its quality depends on other equally important aspects. In addition to psychometric adequacy, the quality of a measure should be evaluated based on alignment between the theoretical and measured constructs (construct validity of the measure) and the alignment between the measured construct and reality (ecological validity of the theoretical construct).

Furthermore, as Messick mentioned, validity is not a property of a scale or measure itself; it is the property of the inference derived or decision based on a measurement (see Winne, 2020). Messick (1989) defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 13). Key aspects of Messick's definition include: validity (1) is a unified concept that encompasses all aspects of the accuracy, appropriateness, meaningfulness, and usefulness of a test or assessment tool; (2) is a matter of degree, not an all-or-nothing property; (3) is based on both empirical evidence and theoretical rationales; (4) is concerned with the inferences and actions based on test scores, not just the scores themselves; and (5) includes the consequences of test use, both intended and unintended (i.e., consequential validity).

Similarly, Furr and Bacharach (2014) defined validity as the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. A measure can be

judged to be valid and reliable enough for a given purpose or given context but judged invalid and not-reliable for other purposes and/or contexts. This means that validity as an integrative concept is highly context-sensitive, and its generalizability requires both empirical evidence and theoretical rationale.

Psychometric adequacy is only useful when construct validity of the measure and the ecological validity of the theoretical constructs are established. It should not be the other way around, where psychometric adequacy of a scale is treated as evidence of overall validity. Beyond the subtypes of validity discussed, the validity of inferences/decisions based on a measurement also depends on factors that might fall outside the realm of measurement.

In summary, while the concepts of measurement error and measurement quality are broader than statistical error and psychometric adequacy, the concept of validity is also broader, with psychometric adequacy being just one of its constituent subtypes. Using the term “validity” to refer to a subtype (e.g., ecological, psychometric) without specifying what is being validated (e.g., validity of theoretical construct, operationalized construct) is restrictive and problematic, as validity is a unified concept at a higher-level compared to its constituents’ subtypes.

What has been done in the studies of this dissertation is mostly at the level of the psychometric validity. Table 7 summarizes the meaning of psychometric validity and statistical error according to the analytic methods used in each study.

Table 7*The Psychometric Validity and Statistical Error Across the Analytic Methods Used*

Analytic Methods	Psychometric Validity of the Measured Construct	Statistical Error
IRT (Study 2)	the degree to which evidence supports the accuracy of interpreting a measure as a representation of the latent attributes it aims to assess (the probability of categorical item responses given the respondent's latent level and the characteristics of the items).	the extent to which the observed item responses deviate from the expected item responses based on the respondent's latent trait level and the characteristics of the items.
CFA, ESEM (Study 1 & 3)	the degree to which evidence supports the accuracy of interpreting a measure as a representation of the latent constructs it aims to assess (the alignment of the hypothesized measurement model to the observed data).	the extent to which the observed indicator variables deviate from their predicted values based on the latent constructs they are intended to measure (the portion of the variance in the observed indicators that cannot be explained by the latent constructs).
LPA (Study 3 addendum)	the degree to which evidence supports accuracy of interpreting a probability that a given set of observed responses belongs to a particular latent profile (conditional probabilities of observed indicators given latent class membership and the characteristics of the indicators).	the extent to which observed indicator patterns deviate from the latent profiles identified by the model (inaccuracies in the classification of individuals into latent profiles).

The bidirectional paths between the theoretical construct and the measured construct in Figure 19 highlight that measurement should be viewed as a continuous and reciprocally determined process that constantly interacts with and shapes the theoretical constructs. This process involves a reciprocal feedback loop, where the measure informs the theory, and the theory, in turn, guides the refinement of the measure. This ongoing interaction ensures that both the theory and its operationalization undergo iterative refinement and improvement.

As shown, none of these psychometric investigations can directly inform how well academic challenges as a theoretical construct represents the students' academic challenges as

perceived in the real world. Similarly, they cannot directly indicate alignment between the theoretical construct of academic challenges (i.e., conceptual definition) and what has been operationalized in the measure as its proxy (i.e., the measured construct).

In sum, while this study examined the psychometric properties of the Academic Challenges scale, the validity of interpretations derived from the scale is contingent on additional layers depicted in Figure 19. The following section briefly exemplifies theoretical considerations and associated uncertainties at these layers. It is important to note that these claims, though discussed in the context of the studies conducted, are primarily informed by theoretical perspectives, and should not be viewed as empirical conclusions originating in the reported findings.

Theoretical Considerations

Challenges as Psychological Situations

The Cognitive-Affective Personality System (CAPS) theory, developed by Mischel and Shoda (1995), proposes that behavior is a function of the person in the situation, emphasizing the importance of *psychological situations*. In other words, behavior results from the dynamic interplay between an individual's personal characteristics (such as their personality, beliefs, and emotions) and their perceptions of the situation (Mischel & Shoda, 1998). Psychological situations are not just physical environments, but how individuals perceive and interpret their circumstances (Shoda et al., 2002). Using a concept called *If-Then Signatures*, this theory posits that while a person's behavior varies across different situations, it shows higher consistent patterns across similar psychological situations (Mischel & Shoda, 1995; Shoda et al., 1994).

Mediation between situation and behaviour occurs through Cognitive-Affective Units (CAUs) in CAPS (Mischel & Shoda, 1995). Based on CAPS, these cognitive-affective units

include: (a) how an individual perceives and interprets a situation; (b) what an individual expects to happen in the situation, including their beliefs about their own abilities and the likely outcomes of different actions; (c) what an individual aims to achieve in the situation (goals) and the importance they place on different outcomes (standards, values); (d) their affective responses; and (e) competencies and self-regulatory skills (Mischel & Shoda, 1998; Shoda et al., 2002).

Academic Challenges can be viewed as reflections of the psychological situations students perceive as they regulate their learning. In other words, they can represent important aspects of the academic context from the student's perspective. The other three scales in SRL-PSD also aim to situate students self-reported behaviors within the context of specific cognitive-affective units, such as standards and self-efficacy beliefs (Jarvela, et al., under review). The *Academic Challenge* profile as an abstract, multidimensional perceived context has potential for creating a backdrop for interpreting and understanding students' behaviors. However, such a conceptualization of academic challenges has certain limitations, which will be discussed in the following section.

Academic Challenges from Reality to the Measured Construct

The conceptualization of academic challenges as reflections of perceived challenges in underlying domains (such as motivation regulation) represents an abstraction from students' concrete experiences. As this theoretical model categorizes specific experiences into broader theoretical constructs, it is important to consider how well the constructs align with students' own perceptions. For example, when a student expresses “I cannot figure out how to tackle this task”, they may not consciously identify this as a metacognitive challenge. The classification of such experiences into a theoretical domain like metacognition is an interpretative step taken by

researchers. Any potential discrepancy between real experiences and an abstraction of the experiences as a construct highlights a critical issue in ensuring the ecological validity of the construct aiming to measure subjective experiences.

Additionally, the alignment between students' perceptions and challenges as theoretical constructs may vary across different constructs. Motivational, volitional, social, and emotional challenges seem to correspond more closely with students' intuitive understanding of their experiences. However, unlike these domains, metacognition may not feature prominently in people's everyday understanding of their behavior. While researchers have developed a rich literature on metacognitive processes and feelings (Efklides, 2006, 2008; Flavell, 1976, 1979; Winne, 2017), students may not naturally distinguish these from cognitive issues. From a student's perspective, experiences like “not knowing how to do something” and “not understanding” might be perceived as similar cognitive challenges. That is, the theoretical model's classification of challenges may, to some extent, deviate from students' lived experiences and perceptions. Consequently, potential misalignments between students' perceptions and the theoretical model's categorization of challenges should be considered when ecological validity of academic challenges is a priority. That is why in Chapter 1, academic challenges were defined as students' perceived challenges, categorized by the theoretical model, into broad regulatory domains. In other words, they are perceptions categorized, not perceptions themselves.

On a related note, the operationalization of academic challenges as correlated domains of a construct assumes that these domains can form an integrated construct with bidirectional associations between domains appropriately represented by covariances. However, this approach may oversimplify the complex relationships between different challenge domains. In reality, some domains may be viewed as outcomes of others, necessitating theoretical, directional

associations to accurately represent their relationships. The nature of associations between domains may vary across different academic challenges, complicating their integration into a single construct. For instance, time management stands out as a unique category. Unlike other domains that primarily involve regulating internal psychological states, time management focuses on managing an external resource, requiring the coordination of various internal processes (Wolters & Brady, 2020). Similarly, procrastination might stem from various self-regulatory issues (e.g., cognitive, motivational, metacognitive, etc.), suggesting a hierarchical relationship in which the regulation of overt behaviors and resources may operate at a different level compared to the management of internal psychological states (Steel & Klingsieck, 2016). Cognitive challenges present another example of this complexity. They can be influenced by factors such as self-efficacy, where a student's perceived learning capability (often considered a motivational belief) can affect their perception of a task's difficulty, irrespective of its complexity (Bandura et al., 1999; Putarek & Pavlin-Bernardić, 2020; Zimmerman, 2000).

These examples highlight the potential limitations of treating all academic challenge domains as interchangeable components of a single construct. A more nuanced approach, considering their varying levels of analysis and potential causal relationships between domains, may be necessary for a comprehensive understanding of academic challenges. The complexity of the associations among these factors and the potential gap between theoretical constructs and the conceptual definition discussed earlier raise important questions about the construct validity of the academic challenges, which will be addressed in the next section.

Pragmatic Considerations

The categorization of students' perceived challenges was based on the major regulatory domains that likely distinguished those instances, with each instance being assigned to a domain

that best reflects its nature. For example, the statement “I struggled with understanding learning materials and concepts” was categorized as a cognitive challenge. Such a multidimensional structure was then examined through empirical methods, particularly factor analysis. The initial development of the measure involved categorizing the list of challenges students reported in an open-ended survey. Therefore, the categorization process can be considered partly data-driven, as it was not entirely based on an *a priori* theoretical framework. Such data-driven models identify patterns and clusters of perceived experiences without necessarily explaining the underlying psychological processes that give rise to them or how they interact. In other words, the theory underlying these constructs is primarily descriptive rather than explanatory. Consequently, the value of translating student perceptions into theoretical categories lies in its potential practical implications such as enhancing student strategy deployment and helping educators improve intervention strategies for the challenges students actually perceive. Put differently, such a theoretical model should be primarily valued from a pragmatic and instrumentalist perspective. Such a view prioritizes the practical applications and usefulness of concepts and principles in producing tangible outcomes, rather than solely focusing on their power in explaining the underlying causal mechanisms of human experiences and behaviours (see Yarkoni & Westfall, 2017, for a similar discussion of predictive vs. explanatory models in psychology research).

To support the theoretical construct of academic challenges, utility checks can be employed in various ways. One example is examining whether raising and updating students' metacognitive knowledge about academic challenges can enable them to better understand and regulate their own behaviors by strategically responding to the challenges they face. Similarly, investigating how effectively the abstract context provided by the academic challenges profile

can guide the selection and implementation of interventions is crucial. This ensures that the measure reflects practically significant areas for change rather than merely non-significant symptoms. To demonstrate how strategy deployment and support provision can be systematized according to the measured construct of academic challenges, both theoretical rationale and empirical evidence are necessary.

Organizing and Tailoring Support Based on Academic Challenges

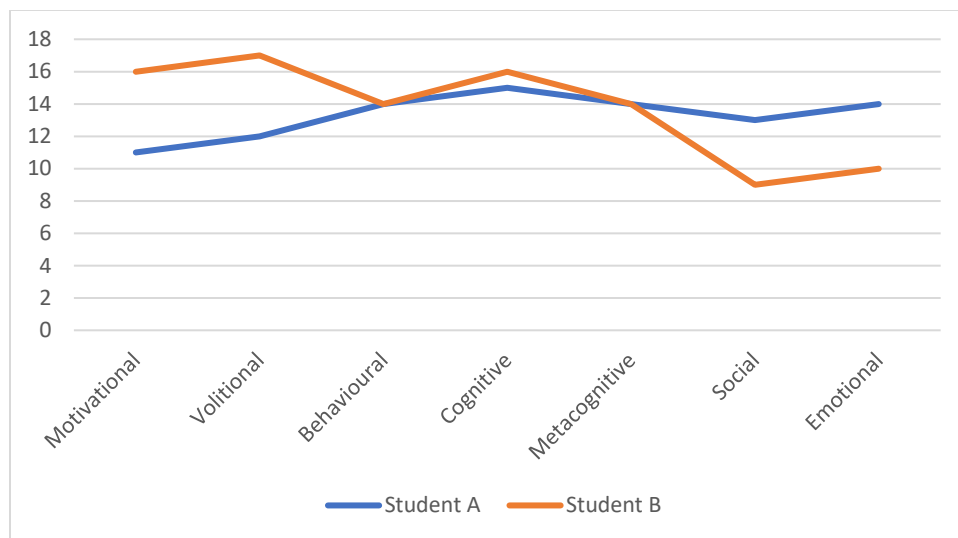
A key assumption supporting the utility of measuring academic challenges is that understanding the specific types of challenges students face can inform the selection of strategies to effectively address them. For example, when confronted with motivational challenges that hinder a student's engagement in the learning process, strategies that enhance interest and incentivize participation, such as self-rewarding, may prove beneficial. Cognitive and metacognitive challenges, on the other hand, may necessitate the use of specific strategies, like breaking down complex information into manageable chunks or employing mnemonic devices to facilitate retention. Similarly, time management challenges may require students to focus on cultivating effective planning and organizational skills, such as creating structured study schedules and prioritizing tasks based on importance and urgency. In the case of emotional challenges, students might need to adopt emotional regulation strategies, including stress management. By gaining a comprehensive understanding of the types of challenges students perceive, educators and learners may be in a better position to apply targeted, personalized strategies and practices that address their unique needs.

However, as demonstrated by the findings from studies 1 and 3, academic challenges are highly interrelated. To exemplify, a student's lack of motivation can intensify cognitive challenges by diminishing their engagement and effort in processing information. Conversely,

difficulties in cognitive processing can result in frustration and decreased motivation. Similarly, poor time management skills can elevate stress levels, leading to emotional challenges that further impede a student's ability to focus and learn effectively. Consequently, addressing one challenge can have cascading effects on other areas, underscoring the importance of adopting a comprehensive and relational approach to understanding and supporting students based on their academic challenges (see Figure 20).

Figure 20

An Example of Students' Academic Challenges Profiles



Note. Academic challenges are often complex and multifaceted, requiring solutions that address the interplay between different regulatory domains rather than treating each in isolation. For example, in this hypothetical scenario, although both Students A and B have the same high level of metacognitive and cognitive challenges, these high levels could be better interpreted in relation to their differences in other areas: social and emotional challenges for Student A, and motivational and volitional challenges for Student B.

After creating a learner's academic challenges profile, the selection of areas for intervention should be based on practical solutions that consider the feasibility of change. Additionally, the chosen areas should be selected such that improvements are likely to indirectly address other observed challenges as well. For instance, if a student reports challenges with time management and emotional challenges, educators must judge whether it is more effective to initially implement strategies that help them manage their time efficiently, directly target stress,

or perhaps tackle both simultaneously. A pertinent question in this context is whether more efficient time management can boost a sense of control and alleviate stress, or if the level of stress is so high that it prevents any attempt to improve time management. Prioritizing practical solutions allows educators and students to implement strategies that are (a) effective, (b) sustainable, and (c) likely to have a cascading effect on other challenges.

To further enhance the utility of the academic challenges scale, future studies should focus on developing a qualitative diagnostic schema that helps in understanding different profiles of academic challenges. Table 8 summarizes factors that potentially impact tailoring support based on measuring academic challenges. What has been discussed is just a preliminary discussion underscoring the necessity of establishing the practical utility of the *Academic Challenges* scale. The effectiveness of these proposed suggestions for tailoring support based on academic challenges requires further theoretical and empirical examination through future research studies.

Table 8

Factors Influencing Tailoring Support Based on Academic Challenges

Contextual Relevance	• The effectiveness of strategies in addressing academic challenges is contingent upon their alignment with the specific nature of the challenges they aim to target. • Different types of challenges may necessitate distinct approaches.
Interconnected nature	• Challenges should be interpreted as an integrated profile rather than discrete factors
Prioritization	• Giving priority to addressing challenge areas that require urgent intervention due to their immediate and significant impact on the student's learning and well-being. • Focusing on addressing challenges that can be more readily and effectively tackled using available resources • Promoting behaviour modifications that are likely to contribute to long-term success • Addressing challenges with the potential to alleviate other related challenge areas.

Summary of Findings

Findings from Scale-Level Psychometric Examinations (Studies 1 & 3)

In this section, I discuss findings from studies 1 and 3, in which I examined the multidimensional structure of the *Academic Challenges* scale. In study 1, the multidimensionality of the scale was investigated using Exploratory Factor Analysis (EFA) and a correlated factor Confirmatory Factor Analysis (CFA) model. The substantial correlations observed among the six factors of the scale suggested the potential presence of an overarching factor influencing responses across these dimensions. Failing to account for such a general factor could compromise the discriminant validity of the scale and lead to biased parameter estimates (Asparouhov & Muthén, 2009). In study 3, both higher-order and bifactor measurement models were employed to examine the possibility of modeling a general construct of academic challenges. The bifactor model proved superior as it allowed for the properly disaggregated coexisting general and specific factors, enabling a more nuanced dissection of these constructs.

Study 3 aimed to address another key limitation in the conventional measurement models of the academic challenges scale: the assumption that clusters of challenges are independent of one another. In a CFA model used in study 1, an item such as “struggling with believing I can do my work” is attributed solely to the motivation challenges scale. However, such an item could also reflect non-primary but major influences from other factors, such as metacognitive challenges (e.g., low confidence due to not knowing how to tackle the task strategically). This interdependency suggests that items may not exclusively belong to a single factor but rather exhibit substantial cross-loadings. Forcing items to exclusively load onto their primary factors in a standard CFA model could result in inflated parameter estimates, potentially distorting

subsequent structural path estimates (Mai et al., 2018; Marsh et al., 2014; Morin et al., 2020; Wei et al., 2022; Zhang et al., 2023).

While bifactor and second-order models were employed to address the need for a general factor, an Exploratory Structural Equation Modeling (ESEM) approach was employed to allow for the estimation of cross-loadings in a confirmatory measurement model. By employing a sequential approach (Morin, 2023; Morin et al., 2016; Morin et al., 2020), the impact of addressing these two limitations of Study 1-lack of a general factor and suppressing cross-loadings-was investigated both individually and collectively.

Based on study 3, the bifactor ESEM emerged as the best-fitting model. This determination was based on a comprehensive review of global fit indicators and comparative assessments of local parameter estimates across various models, including first-order and bifactor models, as well as CFA and ESEM models. In a bifactor ESEM model, both cross-loadings and a general factor are modeled, potentially leading to a more realistic representation of the multidimensionality of the academic challenges.

An important finding from examining the bifactor ESEM model was the strong loadings of the metacognitive challenges items on the general factor, accompanied by minimal loadings on the specific factor of metacognitive challenges itself. This phenomenon, often referred to as the vanishing of specific factors (Morin et al., 2020; van Zyl & ten Klooster, 2022), occurs when the general factor accounts for a significant portion of the common variance among items, leaving minimal variance for the specific target factor to explain. This phenomenon results in minimal and at times anomalous patterns of factor loadings on the specific factor.

Consequently, given the dominance of the general factor in capturing the variance attributed to metacognitive challenge items, the specific factor corresponding to metacognitive

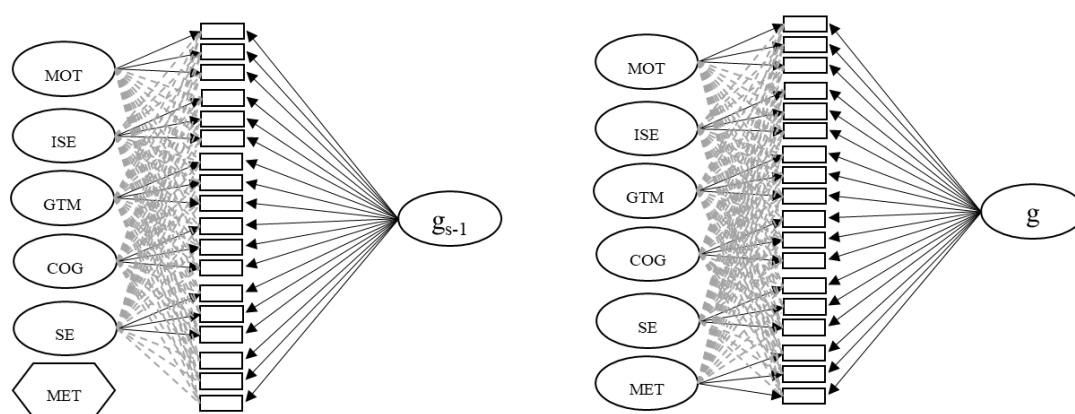
challenges was omitted. This modification led to all eight indicators of this scale being treated as primary indicators of the general factor, adopting a modeling approach known as the S-1 bifactor ESEM model. Given the importance of this decision, it will be discussed in detail in the following section.

Theory-Driven vs. Data-Driven General Factors

In a bifactor model, the general factor (g) represents the shared variance among all indicators, while specific factors (S-factors) capture the unique variance of each subscale beyond what is explained by the g factor. However, the interpretation and stability of the g factor can vary depending on whether a symmetrical bifactor model or a bifactor S-1 model is used (Figure 21).

Figure 21

Symmetrical Bifactor ESEM Vs. S-1 Bifactor Model



Note. In a symmetrical bifactor model (on the right), all specific factors of a scale are modeled, with each item loading on both the general factor and its respective specific factor, and the general factor being defined by all items within the scale. In contrast, in a bifactor S-1 model (on the left), one specific factor (S) is chosen as a reference factor to directly define the general factor (g), and S-1 specific factors are estimated, while other specifications of a regular bifactor model are applied. For clarity, only three items per scale is represented.

In a symmetrical bifactor model, the g factor is defined by all items within the scale, and its meaning must be inferred *post hoc* by examining the pattern of factor loadings. When specific factors differentially contribute to the g factor and are not functionally equivalent, the meaning

of the g factor can become ambiguous and may fluctuate from one study to another (Morin, 2021). Additionally, the variance shared among all indicators as an overarching g factor may include other sources of irrelevant common variance (e.g., common method bias).

According to the stochastic measurement theory, for reflective measurement models to be psychometrically valid, the indicators/domains that the latent factor is assumed to cause should be interchangeable and functionally equivalent (Bollen & Bauldry, 2011; Bollen & Lennox, 1991; Eid et al., 2017; Eid, 2020; Jarvis et al., 2003). In a bifactor model, this means that each specific factor should contribute roughly equally to the overall g factor, without the need to functionally distinguish between the domains. However, when specific factors have causal relationships with each other (e.g., challenges with time management being caused by motivation challenges), this assumption is violated, and the problem is amplified when a whole factor can be plausibly seen as being caused by other factors.

In contrast, in a bifactor S-1 model, one specific factor (S) is chosen as a reference factor to directly define the g factor, and S-1 specific factors are estimated. Choosing the reference factor *a priori* can lead to a more interpretable and replicable g factor (Eid et al., 2017; Eid, 2020). The meaning and interpretation of the g factor in an S-1 model remain consistent as long as the reference domain remains the same, even when other specific factors are added or removed from the model (Haywood et al., 2021; Heinrich et al., 2020; 2023).

The interpretation of S-factors also differs between the two bifactor models. In a symmetrical bifactor model, S-factors capture the variance shared among indicators of a subscale beyond what is explained by the g factor, and their correlations must be fixed to zero for proper disaggregation. In an S-1 model, S-factors represent the part of the factor variance that cannot be explained by the reference factor, and their correlations can be estimated and interpreted (Bureau

et al., 2023; Eid et al., 2017). For example, in Study 3, the motivation scale within the S-1 bifactor ESEM model specifically measures motivational challenges independent of metacognitive influences.

Another practical advantage of the bifactor S-1 model compared to the symmetrical bifactor model is the avoidance of convergence issues in empirical applications. A bifactor S-1 model is a reformulation of a first-order correlated factor model with the same levels of global fit. Because first-order factor models usually do not show anomalous results, the bifactor S-1 model is typically also not affected by estimation problems (Eid et al., 2018).

Choosing the reference factor for an S-1 model should be based on theoretical rationale, not solely on statistical evidence (Burns et al., 2020; Eid, 2020; Eid et al., 2017; Heinrich et al., 2020). In Study 3, metacognitive challenges were chosen as the referent domain as a remedy for psychometric redundancy observed between the g factor and the metacognitive challenges factor in the symmetrical bifactor ESEM model. Findings from Study 2 also showed that the overall total score of the scale predominantly derives its meaning from the metacognitive items. Given the central role of metacognitive processes in self-regulated learning models (Panadero, 2017; Winne, 1996; 2021; Winne & Hadwin, 1998), it was hypothesized that similar to the general factor, metacognitive challenges might arise as a result of the cumulative effect of various other challenges, such as motivational, cognitive, behavioural challenges. When students face multiple challenges in different domains, it can lead to an overall perceived challenges in their ability to effectively monitor and control their learning processes. However, this is merely a hypothesis that requires further theoretical exploration, and its limitations will be discussed in the Future Directions section.

Dominance of the General Factor

The decomposition of variance across the general and specific factors showed that the general S-1 factor accounted for roughly 70% of the common variance in the observed items. This substantial contribution by the general factor underscores its predominant role in explaining the overall variability in the scale's items. Conversely, the specific factors demonstrated suboptimal reliability when the influence of general factor on item variances was partialled out. In summary, the supplementary analyses of reliability and dimensionality (Dueber, 2017; Rodriguez et al., 2016) suggested that although the scale exhibits some degrees of multidimensionality, the existence of a strong general factor may effectively capture the essence of the academic challenges being assessed.

Examining the Criterion Utility of the Measurement Models

Three criterion variables were tested, including students' expected GPA, self-efficacy for GPA, and actual end-of-semester GPA. The analysis of bivariate associations between the factors and criterion variables across different measurement models indicated that a representation of the common variance among observations as a general factor demonstrated consistent associations with criterion variables, irrespective of the measurement modeling approach used. This consistency across models attests to the robustness of the general factor in explaining the items covariances, suggesting that it effectively captures a core aspect of academic challenges that is predictively relevant across various analytical methods.

Further investigation into the bivariate associations of the specific factors revealed that these factors are more strongly associated with the criterion variables in bifactor models where they represent a variance uncontaminated by that of the general factor. Notably, the specific factors exhibited the highest bivariate associations with the criterion variables in the bifactor S-1 ESEM model. The magnitude of these associations was higher than those observed for the

general factor with the same criterion variables, suggesting that, despite their suboptimal reliability, the variance in the specific factors unaccounted for by the general factor represents meaningful constructs rather than merely method effects that can occur in bifactor models.

Additionally, the multivariate analyses which explored associations between criterion constructs and all factors of a measurement model as concurrent predictors revealed that when the general factor was added alongside specific factors as a concurrent predictor of the criterion constructs, the magnitude of associations between specific factors and the criterion constructs were suppressed and became statistically non-significant. This observation suggests that while specific factors were found to be statistically associated with the criterion variables independently, they do not contribute additional explanatory power beyond the general factor in accounting for the variance of the criterion constructs. Although the specific factors do not contribute additional explanatory power beyond the general factor, their modest bivariate associations with the criterion variables support their potential utility as distinct correlates of the constructs that should be further refined. This finding is in line with Study 2 where the specific factors were modelled as independent constructs.

Finally, supporting evidence of measurement invariance demonstrated that the bifactor S-1 ESEM model functions similarly, irrespective of the respondent's sex at birth, their academic year of study, or the cohort (Fall, Winter) to which they belong. The invariance of measurement at the level of latent variance-covariance and latent means supported the feasibility of making comparisons of systematic differences between these demographic groups. Table 9 provides a summary of findings from psychometric examination of the multidimensionality of the measure (Study 1 & 3). Recommendations for using the current version of the scale will be provided in a section titled *Practical Recommendations*.

Table 9*Summary of the Findings from Psychometric Examination of the Multidimensionality of the Measure*

Objective	Strategies tested	Strategies Chosen	Reason(s)
Modeling a General Factor	Second-order Bifactor	Bifactor	• Higher Fit • Properly disaggregation of S and g factors
Modeling Cross Loadings	Correlated CFA Correlated ESEM	Correlated ESEM	• Higher Fit • Improved ecological validity
Modeling a General Factor + Modeling Cross-loadings	Bifactor ESEM, Bifactor S-1 ESEM	Bifactor S-1 ESEM	• Improved ecological validity • Theoretical anchoring of the g factor • Higher Stability and replicability • Removing the psychometric redundancy of the g and metacognitive challenges • Less likelihood of facing statistical anomalies
Objectives	Findings		
Composite Reliability and Dimensionality of the S-1 Bifactor ESEM Model	• approximately 70% of the common variance in the observed items is attributed to the general_{S-1} factor • The scale can be operationalized as essentially unidimensional • Suboptimal reliability for some specific factors (e.g., GTM)		
Bivariate associations between the g and the criterion variables across measurement models	• Various representation of the g demonstrated a consistent association with criterion variables, irrespective of the modeling approach used		
Bivariate associations between the S factors and the criterion variables across measurement models	• the specific factors exhibited the highest associations with the criterion variables in the bifactor S-1 ESEM model • The magnitude of these associations was higher than those observed for the general factor with the same criterion variables		
Multivariate analyses exploring associations between the specific factors and the criterion variables	• Despite strong and significant bivariate associations with the criterion variables, specific factors had limited power in predicting criterion outcome when these variables were regressed on all specific factors at once.		
Multivariate analyses exploring associations of the specific factors and the general factor with the criterion variables	• Specific factors did not provide much incremental validity beyond the general factor in explaining these relationships. • GTM demonstrated incremental validity in predicting criterion validity above and beyond g.		
Measurement invariance	• the bifactor S-1 ESEM measurement model is psychometrically invariant across respondent's sex at birth, their academic year of study, and the cohort to which they belong. • Factor means can be compared across these demographic features.		

Findings from the Complementary Person-centered Analysis (Study 3)

A complementary person-centered analysis was appended to Study 3, which involved comparing two Latent Profile Analysis (LPA) models to further examine the multidimensionality of the *Academic Challenges* scale. The two models compared were: (a) one using scores from the best-fitting first-order model as indicators for the profiles (termed first-order LPA) and (b) one where scores from the best-fitting bifactor model served as indicators for the profiles (termed bifactor LPA).

Results from the first-order LPA revealed that the extracted profiles were characterized mainly by quantitative level differences across the factors. For instance, different profiles exhibited consistently low, medium, or high levels across all challenge areas, indicating a lack of distinct factor-specific patterns that diverged in shape across profiles.

In contrast, in the bifactor LPA, which incorporated both general and specific factors as indicators, level differences among factors were effectively absorbed by the general factor, allowing the specific factors to represent more nuanced shape differentiations.

In line with the results of the main analysis in Study 3, LPA models derived from the symmetrical and S-1 bifactor models also suggested psychometric redundancy of the general and metacognitive factors and supported the use of S-1 bifactor model as a potential solution for the redundancy issue.

However, the findings from the bifactor LPA should be approached with caution due to the low reliability and low factor score determinacy of the specific factors, potentially affecting the robustness of the profiles derived in this analysis.

Despite this, the primary purpose of conducting this clustering analysis was to explore the structural complexities that can threaten the appropriateness of using variable-centered analyses

for modeling the multidimensionality of the construct. For instance, using person-centered analyses as a complementary method can reveal the degree of oversimplification that forcing a variable-centered model may have created onto a sample of potentially heterogeneous groups. Similarly, the existence of complex shape differentiations among the first-order factors, if observed, would pose a challenge to the validity of the general factor as a representation of the covariation of observations and/or factors. This would suggest the need for more complex or even non-linear structures to adequately model what the dimensions of a construct represent in combination. However, the level differentiation among first-order factors in this study indicated that the relationships between indicators could be effectively modeled using variable-person centered analyses, without necessarily resorting to mixture modeling such as latent profile analysis.

Furthermore, there were significant mean differences in terms of the criterion variables across profiles. In the first-order LPA, these differences were observed between profiles that consistently differed in terms of levels of all six challenges, while in the bifactor S-1 LPA, the significant differences were observed between profiles that primarily differed in terms of levels of the general S-1 factor. Overall, these findings underscored the importance of incorporating the general factor to represent the interplay of multiple challenge dimensions into a practical and interpretable structure.

Findings from Item-Level Psychometric Examinations (Study 2)

Item Response Theory (IRT), specifically Samejima's Graded Response Model (GRM; 1997, 2016), was employed in Study 3 as an alternative analytical method to triangulate the scale-level psychometric properties of the academic challenges scale. Unlike the factor analytic approaches used in Studies 1 and 3, which assessed the overall fit of a multidimensional

theoretical model to the observed data, IRT focused on the individual item level. This distinction enabled IRT to complement CFA by providing a detailed evaluation of each item within the academic challenges scale, offering a more comprehensive understanding of key person and item parameters such as the range of latent attributes, discriminations, thresholds, information, and reliability.

Additionally, while CFA presupposes consistent psychometric properties across all levels of the latent attributes, IRT assumes that measurement precision may vary based on the individual's level of the latent attribute being measured (Embretson & Reise, 2013). This characteristic of IRT is particularly relevant to the scale given its diagnostic aims, as it is crucial to understand the range of academic challenges it can reliably measure. Furthermore, IRT allowed for evaluating the effectiveness of the Likert response scale in distinguishing among individuals (Ostini & Nering, 2006) as well as converting theta scores to observed expected scores to provide norm-based reference distributions.

A polytomous unidimensional IRT approach was chosen as the analytic method because it facilitated connecting information regarding the psychometric properties of the scales to the properties of arithmetic scores for the scales that can be easily computed for practical uses. While this approach did not consider the multidimensionality due to specific academic challenges factors in the unidimensional construct of general academic challenges (AC-g), local dependency (high residual covariance between items of a latent factor) was examined for the AC-g scale. The findings supported the stability of unidimensional IRT parameters even in the presence of certain degrees of local dependencies attributed to unmodeled multidimensionality.

The foundational assumptions of IRT, including unidimensionality, local independence, monotonicity, and Differential Item Functioning (DIF), for all specific scales and the overall

academic challenges scale (AC-g) was examined. Internal consistency of the scales revealed adequate reliability for the unidimensional constructs, including the AC-g scale. The analysis indicated the appropriateness of the unidimensional GRM approach for each scale. Finally, GRM configurations with either constrained or unconstrained discrimination parameters were compared, indicating that models with unconstrained discrimination parameters were more suitable.

The findings from the item-level analysis of the *Academic Challenges* scale indicated the measure's adequate psychometric properties both as independent scales and as a unidimensional scale for assessing general academic challenges. However, recognizing scale development and validation as an ongoing and iterative process, recommendations for improvements based on this study will be proposed.

Specifically, two items, SE2 and ISE2, were identified as problematic within both their respective specific scales and the general academic challenges scale. Suggestions on refining the wording of these items and/or addition of new items will be provided later in a separate section. The *initiating and sustaining engagement* scale was identified as the most frequently encountered challenge among students, as evidenced by high agreement rates and low threshold parameters. It was suggested that this scale could be refined by incorporating items that capture more intense levels of challenges related to initiating tasks and managing attention. Similarly, the addition of items to minimize item overlap and to align the *goal and time management* scale more closely with measuring the behavioral and volitional challenges was recommended. Although the current version of the scale demonstrates adequate psychometric properties based on IRT analysis, it was suggested that the social and emotional aspects within the social and

emotional challenges scale could be more effectively measured by dividing them into two separate dimensions.

Adequacy of Composite Scores for Practical Use of the Measure

In addition to being a measurement tool in research, the Academic Challenge scale is designed to serve as a self-assessment and diagnostic tool. Although the scale was initially validated as a 6-factor correlated model through confirmatory factor analysis (Study 1), in practical applications such as in digital self-assessment reports and for classroom usage, sum or raw average scores are often employed to operationalize the scales. As the psychometric adequacy of a model derived from analytic techniques such as CFA, which directly addresses statistical error, does not directly extend to its composite scores (McNeish & Wolf, 2020), examining the adequacy of composite scores (such as summing or averaging) for the practical use of the measure was essential. Consequently, exploring IRT parameters was complemented by examining the adequacy of total scores for both academic challenge factors and the overall scale. Additionally, associations of the scales and their composite scores with three criterion constructs were investigated to provide preliminary information about their practical utility.

Findings indicated that standardization or transformation of scores before aggregation was unnecessary due to the homogeneity of variances across items and the absence of significant univariate normality violations. The evidence of unidimensionality and internal consistency reinforced the suitability of composite scores for both specific scales and the AC-g. Notably, average scores demonstrated a very high correlation with three types of estimated factor scores in this study (.96 to .99), revealing that averaging individual items to compute a total score yielded a stronger correlation with factor scores than averaging the means of specific factors (.979 vs. .967). The study also explored the congruence between true factors and their estimated factor

scores. It was found that the factor scores for *motivation* and *initiating and sustaining engagement* challenges were aligned less closely with their true factors compared to other scales, whereas *goal and time management* and *social and emotional* scales showed adequate alignment, and *cognitive*, *metacognitive*, and general unidimensional academic challenges scales (AC-g) very strongly aligned with their true factors.

Overall, the AC-g's true factor showed a high association with its average scores (.985), underscoring the effectiveness of using an overall average score for the entire scale. However, the findings also highlighted the need for improvement, particularly in the MOT and ISE scales, where approximately 17% and 15% differences between factor scores and true factors were observed, respectively.

Examining the degree to which AC-g overall scores relate to specific factors revealed a significant alignment between the overall AC-g scores and the metacognitive challenges scores. This alignment persisted even after adjusting for the higher number of MET items compared to other scales, underscoring the substantial correlation of metacognitive challenges with the common variance among other challenge areas. This finding distinguished the metacognitive aspects within the academic challenges framework and suggested further investigation into this relationship.

Criterion Utility of Academic Challenges as Unidimensional Constructs

Examining the criterion utility of the scales showed inverse statistically significant associations between five of the six subscales and students' end of semester GPAs. All scales also showed statistically significant correlations with GPA expectations and GPA Self-Efficacy. The social and emotional scale was the only scale, which was not statistically associated with actual GPA, although it was statistically related to GPA expectations and GPA Self-Efficacy.

This unexpected result may be explained by the unprecedented context of the COVID-19 pandemic, restricted social activities, and the transition to online learning modes, which potentially disrupted the expected relationship between social and emotional factors and academic performance. Nonetheless, while this explanation seems plausible, it is important to note that it remains speculative at this point and requires further empirical investigation to be substantiated. The overall scale score (AC-g), indicative of general academic challenges, also demonstrated inverse statistically detectable associations with all criterion variables. Although the magnitudes of these associations were modest, the findings highlight the potential utility of academic challenges in academic regulation and performance.

Suggestions For Improving the Scale

Suggestions for Improving the Measure based on Item-Level Analyses

In study 2, the psychometric properties of the Academic Challenges scale were examined using a set of polytomous graded response IRT analyses. These analyses offered detailed information about individual items and provided insights for item refinement summarized in this section. In this study, each academic challenge factor was initially treated as a unidimensional construct and then items from all factors were combined to form a unidimensional overall construct of academic challenges.

The suggestions for improving the scales presented in this study should not be interpreted as the inadequacy of psychometric properties of the current scale. As mentioned, I stress that the measurement validation process should not end once adequate psychometric indices are achieved; instead, continuous refinement between the theoretical framework and measurement models is essential.

General Suggestions for Improving the Measure Based on IRT Findings

Response Scale. The 5-point response scale, ranging from *strongly disagree* to *strongly agree* with *neutral* at its midpoint, was generally effective for the scales. However, for some items, such as ISE2 and SE2, this response scale was observed to potentially increase cognitive load for respondents. Moreover, the *neutral* response option covered a restricted range of the latent attributes and showed overlap with adjacent response options. To address this, a revised scale with the options *never true*, *rarely true*, *sometimes true*, *usually true*, and *always true* is proposed, along with a prompt like “Think about the past 2 weeks when you rate how true each statement was for you. Over the past two weeks, I struggled with...”. This revised response scale more accurately captures the experience of facing challenges, as opposed to the agreement/disagreement format. Its theoretical distribution is more aligned with a continuous spectrum rather than a dichotomous agreement/disagreement model with neutral as a pivot point. This adjustment has already been implemented in data collection and will be analyzed in upcoming research.

Theta Range. Regarding the range of the latent attribute measured by the specific scales and the AC-g, the target range was set between -2 and 2. The *Academic Challenges* scale is intended as a diagnostic tool for the general student populations, aiming to identify variations up to two standard deviations from the mean. While capturing the remaining 5% is important, it falls outside the intended uses of this self-report measure and may require a different approach specifically designed to measure those degrees of deviation. However, to reliably measure within this chosen range ($r > 0.8$ in $\Theta = -2$ to 2), some specific scales, such as ISE and GTM, require item revision or the addition of new items. Suggestions for improving these scales will be provided in dedicated subsections for each scale.

Including Attention Checks. Incorporating at least two quality control questions into the survey would be beneficial for identifying and excluding participants who would not adequately

engage with the survey questions. An example of such a question could be “please respond *Never true* to this item”. Furthermore, using parallel versions of at least two items and assessing the consistency of responses between them can serve as a complementary method for attention checking. Attention checks can enhance the survey's reliability by ensuring that responses are attentive.

Item Randomization. Considering the observation of higher residual covariance among items from the same scale that are sequenced consecutively, it is essential to randomize questions throughout the entire scale. This randomization would reduce the possibility that the observed residual covariances arise from a method effect, such as item sequencing. Table 10 includes a summary of the proposed modifications discussed above.

Table 10

Summary of proposed modifications to the Academic Challenges scale based on IRT analysis findings

Area of Improvement	Current Approach	Proposed Modification	Rationale
Response Scale	5-point scale (strongly disagree to strongly agree with neutral midpoint)	5-point scale: never true, rarely true, sometimes true, usually true, always true with prompt: “Think about the past 2 weeks when you rate how true each statement was for you. Over the past two weeks, I struggled with...”	- Better captures the experience of facing challenges - Aligns with a continuous spectrum rather than dichotomous agreement/disagreement - Reduces cognitive load for some items (e.g., ISE2, SE2)
Theta Range	Target range: -2 to 2 (two standard deviations from the mean)	- Maintain target range of -2 to 2 - Revise or add items to specific scales (e.g., ISE, GTM) to reliably measure within this range ($r > 0.8$)	- Suitable for a diagnostic tool for the general student population - Capturing the remaining 5% may require a different approach
Attention Checks	Not currently implemented	- Include at least two quality control questions (e.g., “please respond <i>Never true</i> to this item”) - Use parallel versions of at least two items to assess response consistency	- Ensure responses are attentive - Identify and exclude participants not adequately engaged with the survey
Item Randomization	Items from the same scale are sequenced consecutively	Randomize questions throughout the entire scale	- Reduce the possibility of residual covariances arising from method effects (e.g., item sequencing)

Suggestions for Refining the Motivation Challenges Scale (MOT) Based on IRT Analysis²

All three items of the scale demonstrated highly consistent and strong abilities to differentiate between individuals with varying degrees of perceived motivation challenges. The items also showed comparable levels of information, contributing equally to the scale's measurement precision. Together, the items covered the distribution of the motivation challenge latent continuum ranging from -2 to 2 on the θ scale. Within this range, the reliability was approximately 0.8, and the scale exhibited minimal measurement error around the average θ levels.

The three items within this scale are intended to measure struggles with three key aspects of motivation: self-efficacy (MOT1), value (MOT2), and persistence (MOT3). Given the current structure of the latent model, which includes only three items, this model is at the minimum threshold for statistical identification (Kline, 2015). To enable a more detailed understanding of the constructs being measured using latent structure modeling, I propose adding at least one more item to enhance the scale's ability to detect subtle differences in various facets of self-efficacy, value, and persistence, and to provide a more stable and reliable basis for the statistical latent model. Furthermore, the overall internal consistency of this scale was below the .8 threshold and could be improved by adding more items.

² The primary objective of conducting an item-level analysis is to prioritize precision, which inevitably involves discussing the details. The next 6 subsections include suggestions for item-level refinement of the six specific scales (MOT, ISE, GTM, COG, MET, and SE). The discussion is structured in a manner that allows readers who are not interested in these item-level details and are more focused on broader discussions to skip this part without any disruption in reading continuity. Each of these sections, titled "Suggestions for refining the MOT, ISE, GTM, COG, MET, and SE challenges," concludes with a summary table to facilitate comprehension and provide a concise overview of the key points discussed.

For instance, the concept of value within the context of academic engagement is multidimensional, encompassing not only an interest in the task itself but also the perceived benefits of completing the task (i.e., expected reward; Rosenzweig et al., 2019). To improve the scale's psychometric validity and reliability, items that specifically measure both these aspects of value could be included. For example, adding items like “struggled with maintaining a positive view about the benefits of studying/completing learning tasks” would address the value of the task's outcome. When combined with the existing item (MOT2) that assesses the value attributed to the task itself, this would provide a more comprehensive evaluation of the value component. Additionally, measuring the interest in the tasks themselves through items such as “struggled with finding enough enjoyment in studying/doing a learning task” can capture intrinsic motivation. The inclusion of these items would offer a balanced measure of both intrinsic and extrinsic motivational elements that influence a learner's motivation for a task (see Ryan & Deci, 2000). Table 11 includes a summary of the suggested improvement.

Table 11

Summary of Suggestions for Improving Motivation Challenges Scale Based on IRT Findings

Dimension	Suggested Improvement	Rationale
Overall	Increase the number of items to provide a more nuanced understanding of the constructs and improve the statistical stability of the latent model. This will improve the scale's capacity to discern subtle variations in different aspects of self-efficacy, value, and persistence. An example is detailed below.	Add degrees of freedom to the estimation of the latent factor. By incorporating additional items, the overall internal consistency (e.g., McDonald's ω) of the scale can be enhanced, potentially exceeding the .8 threshold. This balanced approach improves the MOT scale's psychometric validity, reliability, and ability to capture the multifaceted nature of motivation.
Value	Add items that distinctly measure both interest in the task and perceived benefits derived from completing the task: 1. “Struggled with maintaining a positive view about the benefits of studying/completing learning tasks” 2. “Struggled with finding enough enjoyment in studying/doing a learning task”	Provides a fuller evaluation of the value component by capturing both intrinsic and extrinsic motivational factors.

Suggestions for Refining the Initiating and Sustaining Engagement Scale (ISE) Based on IRT Analysis

The analysis revealed that all items, except for ISE2, demonstrated high capabilities in differentiating between various levels of the ISE challenge. ISE4 provided the highest level of information, followed by ISE3 and ISE1. Conversely, ISE2 exhibited the lowest level of information and showed lower sensitivity to changes in the latent continuum (θ), suggesting an area for potential improvement. Together, the items provided reliable information for measuring the ISE challenge for participants scoring between -2 and 1 on the θ scale. The scale's reliability remained high at around 0.8 within this range but decreased to 0.6 at a theta value of 2, with a further decline as the theta values deviated from this range.

The high rates of agreement with ISE items and the observation that the scale's reliability was suboptimal for θ values above 1 suggested that participants generally found the ISE items easy to agree with, emphasizing the need for item modification or the addition of items measuring more challenging aspects of initiating and sustaining engagement.

ISE1 and ISE2 assess struggles with initiating tasks, while ISE3 and ISE4 measure challenges related to distractibility. Considering the IRT findings, ISE2 and ISE1 emerged as candidates for refinement due to their suboptimal IRT parameters and the clarity of their phrasing.

The content of item ISE2 (struggled with *putting things off until the last minute (procrastination)*), when combined with response options ranging from *strongly disagree* to *strongly agree*, may present several challenges for accurate measurement. Respondents must interpret the item through a negative construct (“struggled with”) applied to a negative behavior (“putting things off”), alongside response options that include negative choices (e.g., “Strongly

disagree”). This semantic configuration may increase cognitive load, complicating the process for respondents to accurately reflect their experiences and potentially affecting the item's validity.

Furthermore, unlike other items in the scales where disagreement implied avoiding adaptive behavior, disagreement with ISE2 suggested engaging in adaptive behavior (not procrastinating). Additionally, the phrase “until the last minute” might predispose respondents to *agree* with the item, given the high prevalence of procrastination in academic settings (Steel, 2007; Steel & Klingsieck, 2016), further complicating the item's effectiveness.

A more straightforward approach to wording this item would be “I struggled with tackling/overcoming procrastination”. This revised phrasing could potentially enhance the interpretability and reliability of the item. Removing “until the last minute” could shift the focus from a more deliberate and active form of procrastination commonly observed in university settings (Choi & Moran, 2009; Chun Chu & Choi, 2005) to broader procrastination tendencies, including more dysfunctional ones.

The existing item ISE1, “struggled with wanting to do my schoolwork”, might benefit from a slight modification to “struggled with *starting* to do my schoolwork”. This change shifts the focus from a more general desire or interest in schoolwork to the specific action of initiating the task. It more accurately addresses the element of task initiation, removing the confounding factor of direct interest measured as part of the motivation scale. Table 12 summarizes the suggested improvements.

Table 12*Suggestions for Improving ISE Items Based on IRT Findings*

Item	Current Item	Suggested Improvement	Rationale
Overall	-	Add/modify items to measure more challenging aspects of initiating and sustaining engagement.	High rates of agreement with ISE items and suboptimal reliability for Θ values above 1 indicate that participants found the ISE items easy to agree with, highlighting a need for item modification or additional items.
ISE2	Struggled with putting things off until the last minute (procrastination)	Struggled with tackling/overcoming procrastination	Simplifies the wording to reduce cognitive load, avoids negative construct applied to negative behavior (with new response option recommended earlier), and shifts focus to broader procrastination tendencies.
ISE1	Struggled with wanting to do my schoolwork	Struggled with starting to do my schoolwork	Focuses on task initiation more accurately, removing the confounding factor of interest measured as part of the motivation scale, enhancing the clarity and relevance of the item.

Suggestions for Refining the Goal and Time Management Challenges Scale (GTM) Based on IRT Analysis

All items showed high levels of discrimination, with varying degrees of information provided. Item GTM2 exhibited the highest level of information, followed by GTM4 and GTM1, while item GTM3 displayed the lowest level of information and sensitivity to changes in the latent continuum of GTM challenge.

The items effectively covered a range of the GTM challenge latent continuum from -2 to 1.8, with reliable information mostly obtained at an average Θ value. The scale's reliability remained slightly above 0.8 for Θ levels within this range but decreased to 0.7 at a theta value of 2 and continued to decline as Θ values extended beyond this point.

When considering item GTM3 (struggled with meeting deadlines), which showed the lowest information among the GTM challenge items, it is crucial to recognize the potential impact of contextual factors, such as the timing of data collection. If data is collected during the early weeks of a semester, students might not have encountered enough instances requiring them

to meet deadlines, potentially skewing their responses. To address this limitation, a broader item capturing a wider range of time management issues, such as “struggled with finishing tasks on time” or “within the time frame I planned”, could be more universally applicable.

To obtain a more accurate assessment of students' time management challenges, additional questions in the survey's demographic section could be beneficial. These questions could ask students to rate various aspects of their academic workload, such as the phase of the semester, overall schoolwork load, hours dedicated to studying, and the number of courses taken, etc. This information would provide valuable context and allow for a more nuanced understanding of each student's situation, enabling the treatment of varying academic demands as moderators in the analyses.

Another potential area for improvement is to reduce content overlap between items. The current items, particularly GTM2 (struggling with managing my time, tasks, or goals) and GTM4 (struggling with organizing time and tasks), demonstrate some overlap, as evidenced by their high residual covariance. The GTM dimension is intended to measure challenges within the behavioral aspects of regulation. However, time management is a complex behavior influenced by various facets of regulation, including metacognition and motivation (Wolters & Brady, 2020). Goal and time management challenges can arise due to struggles with planning, which is largely a metacognitive issue involving difficulties such as poor prioritization and low task understanding (Bahena-Olivares et al., 2024). An example item measuring this aspect is “struggled with planning and scheduling”, capturing the difficulties in translating intentions to plans. Another potential cause of GTM challenges is problems with translating plans into actions (behavioral and volitional aspect). Translating intentions to plans is the phase where goals meet plans, while translating plans into actions is the phase where plans meet execution (Higgins, 1998; 2014; Higgins et al., 2003; Jaramillo & Spector, 2004; Jostmann & Koole, 2008; Kazén et al., 2008; Kuhl & Beckmann, 1994; Kuhl et al., 2020). This aspect can be measured using items like “struggled with carrying out what I planned”, reflecting difficulties in executing plans. Given the existence of separate scales for measuring motivational and metacognitive challenges,

this reconfiguration can create space for the behavioral and volitional dimensions crucial for effective time and task management. The items suggested in Table 13, while not exhaustive, aim to showcase examples of items that can capture the behavioral and volitional challenges associated with managing tasks and time.

Table 13

Suggestions for Improving GTM Items Based on IRT Findings

Dimension	Current Items	Suggestions/Alternative Items	Rationale
Overall	-	- Include demographic questions to assess various aspects of academic workload, such as the phase of the semester, the overall schoolwork load for the past two weeks, the number of hours dedicated to studying, and the number of courses taken, etc. - Reduce content overlap between items, particularly GTM2 and GTM4. - Avoid double barreled items - Include items that capture difficulties in translating plans into actions rather than in translating intentions to goals	Provides valuable context and allows for a more nuanced understanding of each student's situation and possible moderators of the construct. Improves distinctiveness and reduces method effects and redundancy in the scale items. Improve accuracy of responses Captures the behavioral/volitional challenges in time and task management rather than metacognitive challenges.
Commitments to Plans	GTM1. Keeping my commitments or goals	Struggled with - staying committed to my plans - sticking to my planned schedule - carrying out what I planned - regularly working towards my goals - establishing study habits (routines) that help follow my plans	Captures various aspects of commitment and behavioural enactment, addressing both behavioral and routine challenges.
Time Management	GTM2. Managing my time, tasks, or goals GTM4. Organizing time and tasks	Struggled with - planning and scheduling - creating a schedule that I can stick to - scheduling and allocating my time	Addresses both planning and execution challenges in time management, reducing content overlap and providing clearer distinctions between items.
Deadlines	GTM3. Meeting deadlines	Struggled with - finishing tasks within the time frame I planned - finishing tasks on time - finishing tasks before deadlines	Broadens the scope to include self-imposed deadlines and makes the item more universally applicable.

Suggestions for Refining the Cognitive Challenges Scale (COG) Based on IRT Analysis

All six items on the COG challenges scale demonstrated high to very high capabilities in differentiating between individuals with varying levels of perceived cognitive challenge.

COG2 provided the highest level of information, followed by COG1, COG4, and COG6. In contrast, items COG3 and COG5 exhibited the lowest levels of information, suggesting a reduced sensitivity to changes in the latent continuum of cognitive challenges. The combination of items covered the distribution of the cognitive challenge latent continuum, ranging from -2 to 2. The scale showed a relatively high reliability, with a coefficient of .85, and maintained low measurement error within this range.

The lower measurement precision of items COG5 and COG3 offered insights for refining the scale. For item COG5 (“struggled with remembering things I learned”), students generally found this item easier to agree with compared to other items on the scale. Despite this, the item captures a crucial aspect of cognitive regulation – the ability to recall learned information. While it may not discriminate as effectively among those with higher perceptions of cognitive challenge, it remains an important element of the scale. To better differentiate the severity of memory challenges among respondents, an alternative item measuring a more severe challenge could be “I struggled with remembering things I learned even after reviewing them”.

Item COG3 (struggling with language and communication) presents a more complex scenario, encompassing two distinct components: struggling with language and struggling with communication. The communication aspect could be interpreted as difficulties in various cognitive processes, such as expressing learned concepts, communicating understanding, sharing knowledge clearly, translating knowledge into understandable language, and discussing content coherently. However, given the sample comprised both domestic and international students,

“struggling with language” might have been primarily interpreted as challenges faced while studying in a foreign language by international students. Separating these components into distinct items might not be practical, as responses to potential items like “struggled with communicating my understanding of the concepts” or “struggled with translating my knowledge into understandable language” also interweave communication and language skills. Nevertheless, these suggested items may be more informative for a scale measuring cognitive challenges because they address more specific and cognitively demanding challenges, which can help correct the positively skewed rates of disagreement in the original item.

Gathering further demographic details, such as the congruence between students' native language and the instruction language, alongside self-assessed language proficiency, can be advantageous. This strategy enables the scale to distinctly identify whether the challenges stem from broad cognitive and communication difficulties or are primarily linked to language proficiency, refining the scale's precision and boosting its accuracy, especially in diverse international student contexts.

Exploring a potential refinement of item COG1 might also be advantageous. Originally phrased as “struggling with applying my learning”, this item is robust in its psychometric properties. Nonetheless, adjusting its wording to “struggled with making learned concepts work in actual tasks” or “struggled with putting my theoretical knowledge into practice” could improve clarity and target a more specific aspect of cognitive challenges. These modifications shift the focus from a broad, general application of learning to a more specific and practical application within the two weeks prior to measurement in an academic setting. Such a narrower focus could provide a clearer assessment of the students' ability to translate theoretical understanding into practice. Table 14 summarizes the suggested improvements for this scale.

Table 14*Suggestions for Improving COG Items Based on IRT Findings*

Dimension	Current Items	Suggestions/Alternative Items	Rationale
Overall	-	Include additional questions in the demographic section to assess potential moderators like phase in semester, number of classes, work-life balance, workload, etc. Also add items to check the congruence between students' native language and the instruction language, alongside self-assessed language proficiency.	Provides valuable context and allows for a more nuanced understanding of each student's situation by taking into account the potential moderators of the construct. Distinguishing between cognitive and language proficiency challenges, especially in diverse international student contexts.
Memory Challenges	COG5. Struggled with remembering things I learned	struggled with remembering things I learned even after reviewing them	Better differentiates the severity of memory challenges among respondents by addressing a more intense aspect of cognitive challenges.
Language and Communication	COG3. Struggling with language and communication	Struggled with - communicating my understanding of the concepts - Struggled with translating my knowledge into understandable language	Makes the items more specific and informative for assessing academic cognitive challenges rather than general communication challenges.
Application of Learning	COG1. Struggling with applying my learning	Struggled with - making learned concepts work in actual tasks - putting my theoretical knowledge into practice	Shifts the focus from a broad, general application of learning to a more specific application, providing a clearer assessment of the students' ability to translate theoretical understanding into practice.

Suggestions for Refining the Metacognitive Challenges Scale (MET) Based on IRT Analysis

All eight items on the MET scale showed very high and consistent capabilities in differentiating between individuals with varying levels of perceived metacognitive challenge.

MET3 and MET6 provided the highest levels of information, followed by MET4, MET2, MET8, and MET1. Items MET5 and MET7 exhibited the lowest levels of information and were less sensitive to changes in the latent continuum of metacognitive challenges compared to the

other six items. Despite this, all items collectively covered the distribution of the metacognitive challenge latent continuum (θ), ranging from -2 to 2. Reliable information was obtained at an average MET challenge continuum, with MET3 and MET6 providing the highest information and MET7 the lowest. The overall reliability of the MET challenge scale remained very high, around 0.95 for θ levels ranging from -1.8 to 1.8, and around 0.9 for theta levels extending from -2 to 2, indicating high reliability and low measurement error within these ranges.

Although MET5 and MET7 showed lower information relative to the other six items in the metacognitive challenge scale, their absolute psychometric properties were still robust. The following suggestions may help refine these items further.

MET5, currently worded as “struggled with prioritizing or figuring out what is important”, could be rephrased as “struggled with prioritizing tasks based on their importance” or “struggled with distinguishing between high and low priority tasks”. The new phrasing avoids the ambiguity introduced by the word “or” and clarifies that the challenge lies not just in the abstract concept of prioritization but in the practical application of assessing task importance and organizing tasks accordingly. This increased specificity can potentially enhance the item's ability to measure an individual's challenges in task prioritization.

The item MET7, “struggled with setting goals for studying”, could be complemented by an item with potential higher threshold parameters such as “struggled with setting realistic and achievable goals for studying”. This modification introduces specificity to the item, better capturing the metacognitive processes involved in effective goal setting. By emphasizing the qualities of goals as realistic and achievable, the revised phrasing may more accurately target the evaluative and regulatory aspects of metacognition, as the item's focus is on the strategic planning skills necessary for successful learning and academic performance. Such specificity

could potentially reduce the cross-loadings between this item and items in the *Goal and Time Management challenges* scale.

Additionally, while less critical, the item MET1 “struggled with breaking tasks down” could be improved to “struggled with breaking tasks down into manageable steps”. This addition further emphasizes the metacognitive skill of breaking larger tasks into smaller, more achievable segments. Table 15 summarizes the suggested improvements for this scale.

Table 15

Suggestions for Improving MET Items Based on IRT Findings

Dimension	Current Items	Suggestions/Alternative Items	Rationale
Task Prioritization	MET5. Struggled with prioritizing or figuring out what is important	Struggled with - prioritizing tasks based on their importance - distinguishing between high and low priority tasks	Avoids the ambiguity of "or" and clarifies that the challenge lies in the practical application of assessing task importance and organizing tasks accordingly, enhancing the item's specificity and measurement accuracy.
Goal Setting	MET7. Struggled with setting goals for studying	- Struggled with setting realistic and achievable goals for studying	Introduces specificity by emphasizing the qualities of goals as realistic and achievable, more accurately targeting the evaluative and regulatory aspects of metacognition and potentially reducing cross-loadings with items in goal and time management challenges.
Task Breakdown	MET1. Struggled with breaking tasks down	- Struggled with breaking tasks down into manageable steps	Emphasizes the metacognitive skill of breaking larger tasks into smaller, more achievable segments, improving the clarity and focus of the item.

Suggestions for Refining the Social and Emotional Challenges Scale (SE) Based on IRT

Analysis

The six items demonstrated a range from moderate to very high capabilities in differentiating between individuals with varying levels of perceived SE challenges. Item SE4 provided the highest level of information, followed by SE6 and SE5, suggesting these items are particularly effective in capturing variations in the latent continuum of SE challenges. In contrast, items SE1, SE3, and SE2 displayed lower levels of information, indicating a reduced

sensitivity to changes in the SE challenge continuum compared to the other items, with SE2 demonstrating the weakest performance.

The items collectively covered the distribution of the SE challenge latent continuum (θ), spanning a range from -2 to 2. The scale was most informative and exhibited the lowest measurement error around the theta level of -0.2. The reliability of the scale remained robust, around 0.9 and 0.8 for theta levels within the -1.8 to 1.1 and -2 to 2 ranges, respectively.

The item SE2, “struggled with feeling lonely”, had a similar issue to what was observed with ISE2. Unlike other items where disagreement suggested that a positive experience, such as feeling connected to others, was not challenging, disagreeing with this item implied that experiencing an undesirable feeling (loneliness) was not challenging. While this might not pose an issue in another context, it is noteworthy that ISE2 and SE2 were unique in having this characteristic, and both showed the lowest precision of measurement across all items. When answering SE2, respondents must navigate a negative construct (struggled with) applied to a negative feeling (loneliness), along with options that include negative responses (e.g., disagree). This setup could heighten cognitive load, making it harder for respondents to accurately convey their experiences.

Analyzing the correlation of the SE challenge items with items in other scales showed that they were less correlated with the other scales compared to items in other scales. This observation suggested that the SE items could be more effectively integrated into the scale if they were tailored to measure academic social and emotional challenges rather than general social and emotional challenges.

Given the title and intent of the scale, which encompasses both socialization and emotional challenges, a reorganization may be beneficial. The scale could be divided into two

distinct facets: socialization challenges and emotional challenges. This separation would allow for a more focused and nuanced measurement of each aspect. The socialization dimension would assess sense of belonging to and interpersonal connections within the academic context (Pedler, et al., 2021; Won et al., 2018, 2019, 2022). This factor could include items such as “struggled with feeling integrated into the academic culture”, “struggled with building meaningful friendships in the academic community”, “struggled with connecting with peers/classmates/instructors at school”, “struggled with establishing a supportive network at school”, “struggled with feeling like I am part of the school/university community”, and “struggled with feeling comfortable and accepted at school”³. These items would collectively gauge the student's level of integration, comfort, and connections in their academic environment.

In addition, what is labeled as an emotional challenges dimension could focus on aspects of wellbeing in the academic context. This can include struggles with managing emotions and feelings, managing stress, feeling frustrated/disappointed at university, being satisfied with one's own academic experiences, taking care of mental health and wellbeing, etc. This dimension could be called academic wellbeing challenges.

By clearly separating these two dimensions, the scale could more precisely capture the complexities of students' experiences in higher education. Identification of the various social and wellbeing challenges students face can facilitate the development of targeted interventions and support systems tailored to students' specific needs to enhance their overall academic experience.

Table 16 summarizes the suggested improvements for this scale.

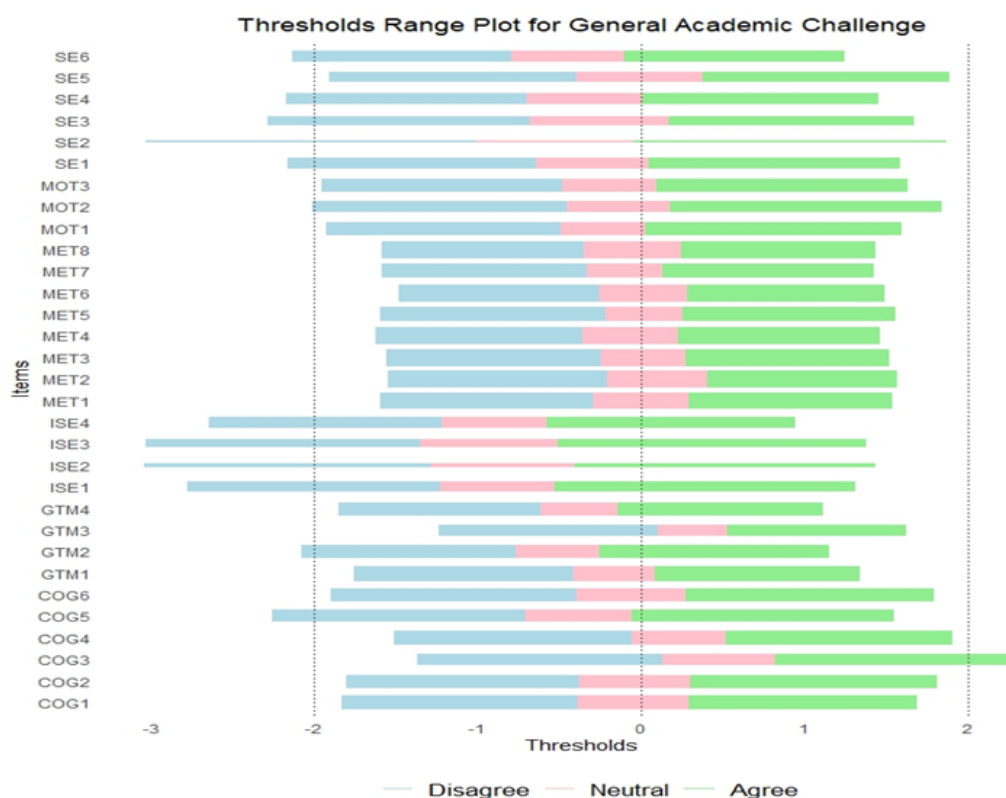
³ “/” was used for a concise combination of examples within a single item. However, it is crucial to separate the suggestions into distinct items to prevent double-barreling. When items merge two or more themes, respondents might concentrate on different themes within the same item, which can increase measurement errors. Separating these themes into individual items ensures clearer interpretation and more accurate data collection.

Table 16*Suggestions for Improving SE Items Based on IRT Findings*

Dimension	Current Items	Suggestions/Alternative Items	Rationale
Overall	-	Divide the scale into two distinct facets: socialization challenges and emotional challenges. Tailor the items to measure academic social and emotional challenges rather than general social and emotional challenges.	This reorganization allows for a more focused and nuanced measurement of each aspect, improving the integration of SE items into the scale.
Socialization Challenges	SE2. Struggled with feeling lonely	Struggled with - feeling integrated into the academic culture - building meaningful friendships in the academic community - connecting with peers/classmates/instructors at school - establishing a supportive network at school - feeling like I am part of the school/university community - feeling comfortable and accepted at school	Addresses socialization challenges by focusing on integration, interpersonal connections, and sense of belonging within the academic context.
Emotional Challenges		Struggled with - managing emotions and feelings - managing stress - feeling frustrated/disappointed at university - being satisfied with one's own academic experiences - taking care of mental health and wellbeing	Focuses on academic wellbeing challenges, emotion regulation, stress, and overall satisfaction with academic experiences, allowing for a more nuanced understanding of students' emotional challenges.

Suggestions for Refining the Unidimensional General Academic Challenges Scale (AC-g)***Based on IRT Analysis***

The main IRT analysis of the general academic challenges scale, which comprises all 31 items, yielded important findings. Generally, students perceived cognitive and metacognitive challenges as less challenging than the other challenge areas, as indicated by the higher rate of disagreement in response to these challenges. *Initiating and Sustaining Engagement* was the most frequent area, as indicated by the highest rates of agreement with the ISE items (Figure 22).

Figure 22*Thresholds Range Plot for AC-g Scale*

Note. The horizontal bars represent threshold ranges for each item. The thickness of these bars signifies their level of discrimination, with thicker bars indicating greater discrimination. The leftward floating position of the bars suggests a high rate of agreement with the items, while a relative position to the right indicates a high rate of disagreement with the items. Vertical dotted lines at $\theta = -2, 0$, and 2 are drawn to facilitate comparisons among the items and to illustrate the latent ranges where the items fail to measure reliably.

The analysis showed that 17 out of the 31 items on the *AC-g* scale provided high information, and 8 items provided moderately high information, indicating their substantial contribution to precisely measuring students' levels of general academic challenge across a -2 to 2 range (Table 17). This analysis also revealed that items from the *Metacognitive Challenges* scale offered the highest level of information to the *AC-g* compared to other scales. In contrast, items ISE2 and SE2 demonstrated very little sensitivity in response to changes in the latent continuum of the *AC-g* challenge.

Table 17*AC-g Items Categorized in terms of IRT Information*

Very High Information (2-2.5)	MET6.Monitoring my task or learning progress MET3.Knowing how to tackle tasks MET8.Applying strategies/techniques properly MET4.Knowing how or when to fix strategies/study skills MET2.Interpreting tasks and expectations MET1.Breaking tasks down
High Information (1.5)	GTM1.Keeping my commitments or goals MET5.Prioritizing or figuring out what is important MET7.Setting goals for studying
Moderately High Information (1)	MOT1.Believing I can do my work MOT3.Persisting when things got tough GTM4.Organizing time and tasks GTM2.Managing my time, tasks, or goals COG1.Applying my learning COG4.Identifying important ideas COG6.Understanding learning materials and concepts COG2.Elaborating and connecting what I learned
Moderate Information (.75)	COG5.Remembering things that I learned MOT2.Feeling like my work was worth doing SE1.Feeling connected SE4.Managing my emotions/feelings SE6.Taking care of my mental health and well-being COG3.Language and communication GTM3.Meeting deadlines ISE4.Focusing or concentrating
Low (.5)	ISE1.Wanting to do my schoolwork ISE3.Managing distractions SE5.Managing relationships SE3.Finding enjoyment at university
Very low information (<.5)	ISE2.Putting things off until the last minute (procrastinating) SE2.Feeling lonely

The quality of the items that provided low or minimal information for the AC-g scale should be evaluated against their performance for their respective specific factors. Where items demonstrated poor quality across both specific and general factors, the recommendations for their refinement outlined in previous sections apply. For example, items ISE2 and SE2 exhibited poor reliability in measuring their corresponding factors as well as in measuring AC-g, indicating that the measurement problem lies with the items themselves. In contrast, items ISE3 and ISE4 provided relatively low information for measuring AC-g but substantial information as items of the ISE subscale. Therefore, their low contribution to the AC-g is likely related to the unmodeled

specific factors in the measurement model of AC-g rather than the quality of the items themselves.

Overall, all items collectively covered the distribution of the general academic challenge latent continuum (θ). The reliability of the scale was around 0.95 for θ levels ranging from -2 to 2 and around .9 for θ levels ranging from -3 to 3. Compared to specific scales, the AC-g had the highest reliability and lowest measurement error across a wider range of θ . Finally, the model predicted that while the scores can range from 31 to 155, approximately 95% of randomly selected participants are likely to score between 49.9 and 142.8, with an average of 98.2.

The investigation into creating a shorter version of the measure revealed that a shorter version cannot reliably measure specific areas of academic challenge. It is suggested that to achieve comprehensive coverage and ensure reliability, each scale should contain at least 4 items, with 5 or more being even better. For instance, the MOT scale, in particular, was noted as needing additional items to enhance its effectiveness. Regarding the overall academic challenges scale, although 25 out of the total 31 items were recognized for their moderately high to very high information value, the development of a shortened scale featuring key items from each specific scale should be deferred. Such a decision should follow the application of the study's recommended modifications on specific scales, allowing for a revised scale that fully reflects an updated understanding of its structure and psychometric strengths. Suggestions for improving AC-g items are provided in Table 18.

Table 18*Suggestions for Improving AC-g Items Based on IRT Findings*

Suggestions/Alternative Items	Rationale
- Apply all previously suggested improvements for the specific scales to the AC-g scale as well. - Include items measuring more intense aspects of initiating and sustaining engagement. - Modify SE2 and ISE2 as they do not contribute effectively to their target specific factors and to the AC-g. - Defer the development of a shortened scale until the recommended modifications have been applied and the revised scale structure and psychometric properties have been reassessed. - Investigate the appropriateness of using the 8 metacognitive challenges items as a strong correlated proxy of the AC-g.	Refinement of items within specific academic challenges domains enhances AC-g precision, relevance, and psychometric strength. ISE item responses were negatively skewed compared to other items, indicating that most students rated their high agreement to these items.

Suggestions for Improving the Scale based on the Scale-Level Psychometric Examinations

While suggestions for item-level refinement were made earlier, the scale could benefit from general improvements, both in its structure and in the analytical methods used to model it, which are detailed in this section.

The study's findings provide several insights that suggest directions for both advancing research on academic challenges and improving the academic challenges scale. Based on the observed dynamics between general and specific factors in the bifactor models, specific methodological adjustments, and extensions to improve the scale's utility will be discussed. These recommendations aim to enhance the accuracy, reliability, and interpretability of the scale by addressing potential issues identified through psychometric analyses of Study 1 and Study 3.

Rethinking Granularity of the Scale

The results highlighted that the specific factors exhibited reduced reliability within the bifactor models due to the high interconnectedness among the challenges and the dominance of the general factor. In other words, the general factor absorbs much of the shared variance that

might otherwise be attributed to specific factors. The predominantly level-differentiated profiles identified in the person-centered analysis suggested that encountering a certain level of challenges in one area is likely correlated with experiencing similar levels of challenges in other areas. Also, while both general and specific factors were found to be associated with criterion constructs as independent constructs, specific factors had limited incremental power in predicting the criterion constructs above and beyond the general factor. These findings invite reconsideration of the granularity of the scale. The scale measuring general academic challenges could potentially be shortened, while the scales for measuring specific challenges should be more specific and targeted. One approach to solve this issue would involve developing a concise scale for assessing overall academic challenges and creating more focused psychometrically independent scales for each specific challenge area. Another approach would involve adding items to the specific scales with the potential to reflect unique aspects of each challenge domain that are likely to remain unaccounted for by the general factor. The first approach seems easier and has potential for enhancing the scales' practical utility and diagnostic capabilities. The second approach allows for modelling both general and specific factors to be reliably represented in a bifactor model. Regardless of the approach taken, refining the granularity of the scale can potentially enhance the reliability and predictive utility of both the general and specific factors.

Prioritizing Ecological Validity in Defining the General Factor

As discussed, observations in a reflective measurement model should be interchangeable and functionally equivalent (Eid et al., 2017; Eid, 2020; Jarvis et al., 2003). Functional equivalence is closely tied to interchangeability, implying that in a reflective measurement model, there should be no major causal relationships between indicators or domains of a construct. The only permitted causal relationships are between latent factors and indicators, not

among latent factors or indicators themselves. Functionally non-equivalent observations might be more aptly modeled using a variant of the formative linear combination approach, where indicators form the composite factor rather than reflect it (Hair et al., 2017). This issue becomes more pronounced when an entire factor can reasonably be viewed as an outcome of other factors. For example, time management challenges are likely influenced by motivational challenges, among other factors.

In the context of study 3, the selection of the metacognitive scale as a referent factor in the bifactor S-1 model was done *post hoc* to solve the psychometric redundancy of this factor and the general factor. The redundancy might indicate that the metacognitive challenges items represent a different functional level of challenges compared to other factors. Similarly, some items from the *Goal and Time Management* (GTM) factor, such as GTM2 and GTM4, may operate at a different level compared to other items within this scale. For example, items like “I struggled with managing time and tasks” might better serve as indicators of general academic challenges rather than a specific factor. In other words, the failure to manage tasks appears to be a better reflection of a pervasive failure across various regulatory domains, rather than reflecting a dysregulation within a single specific regulatory domain.

To address the psychometric redundancy between the general and metacognitive scales in future iterations of the measure, the following modifications are proposed:

1. Avoid including functionally non-equivalent domains within a multidimensional measurement model, particularly one with a hierarchical structure. This is a prerequisite for deriving a data-driven general factor that is interpretable.

2. Treat potentially non-equivalent factors as independent constructs rather than components of a multidimensional construct. The relationships between these factors might be better examined in a structural model instead of a measurement model.

For cases where minor non-equivalence may exist when modeling the g factor, consider:

- a. Defining the general factor *a priori* by incorporating questions that directly measure it (e.g., In the past two weeks, I generally found my academic work challenging). This option has two advantages: (a) the general factor is anchored on what students perceived as the general level of challenges rather than being empirically defined based on their responses across specific indicators of challenges, and (b) this approach would allow for a clearer distinction between general and specific challenges without the need to collapse any of the specific scales (Eid et al., 2017).

While in S-1 bifactor models, the meaning of the general factor is derived mostly based on *a priori* assumptions, and in symmetrical bifactor models through *post-hoc* interpretation, in this approach the general factor reflects students' subjective evaluations of the level of challenges they face. Based on the conceptual model introduced at the beginning of the chapter, the S-1 bifactor model prioritizes the construct validity of the measure, which refers to the alignment between the conceptual definition and the operationalized construct. Consequently, the g_{s-1} factor is an appropriate strategy as long as theoretical rationale supports it. The approach proposed here, however, prioritizes the ecological validity of the measure, as it more closely resembles the subjective experiences of students. A

comparison of these approaches would allow for examining the accuracy of *a priori* assumptions and *post hoc* interpretations made on the g_{s-1} factor.

- b. When there is robust theoretical support, it may be appropriate to anchor a data-driven g factor within a particular domain, as done in study 3. The *Metacognitive Challenge* items exhibited the strongest correlations with g making concurrent modeling of these two factors psychometrically redundant. This was interpreted as suggesting that metacognitive challenges could theoretically stem from challenges across various domains. However, further theoretical investigations are needed to support this decision. Consequently, this empirical finding should not serve as a foundation for future modeling of the scale. Rather, it should inform refinements to subsequent versions, aiming to reduce or eliminate such redundancy.

Restructuring the Measurement Model by Integrating Volitional Processes

The existing scale of *Initiating and Sustaining Engagement* already captures elements of volitional challenges. However, a more explicit focus on volitional processes could potentially enhance the model's explanatory power. By merging this scale with aspects of the *Goal and Time Management* scale—specifically those related to translating plans into actions—a comprehensive *Volitional Challenges* factor could be created. This addition would capture the challenges individuals face in initiating, maintaining, and completing goal-directed behaviors, particularly in the face of competing desires or distractions. Volitional challenges encompass aspects such as procrastination, difficulty in overcoming inertia, struggles with sustained effort, and problems with impulse control. By including this factor, the model would provide a more comprehensive view of task execution (Higgins, 1998;

2014; Higgins et al., 2003; Jaramillo & Spector, 2004; Jostmann & Koole, 2008; Kazén et al., 2008; Kuhl & Beckmann, 1994; Kuhl et al., 2020).

Such a reorganization would not only streamline the model but also provide a clearer distinction between volitional challenges and resource management. Consequently, the *Time Management scale* could be measured as a separate scale distinct from the volitional aspects of task and goal management. This restructuring would potentially improve the model's predictive utility across various domains of performance.

Collecting Data on Potential Temporal, Contextual, and Individual Moderators

Academic challenges are not static but may fluctuate throughout the semester and vary based on individual circumstances. Early in the semester, students may struggle with adjusting to new environments and forming social connections, while pre-exam periods often bring anxiety and cognitive demands. The intensity of these challenges can be moderated by factors such as number of courses one has taken, perceived workload, study hours, work commitments, and overall work-life balance. For example, without this contextual data, a reported low level of academic challenge cannot be differentiated between simply reflecting a light course load or an average load combined with low perceived challenges. Additionally, a student's status as international or domestic may impact their experience with newcomer international students potentially facing additional hurdles related to cultural adjustment and language barriers. First-year students, particularly in their first semester, may encounter unique challenges as they transition to university life. Moreover, the specific major or field of study can introduce distinct academic pressures and requirements.

Additionally, beyond these external contextual characteristics, a student challenge profile can also fluctuate even if the context remains unchanged. This can occur when students adjust

their personal standards, goals, or effort levels, among other factors. Therefore, in addition to the characteristics of external context collecting additional constructs that measure their performance expectations and desired goals is recommended.

To gain a comprehensive understanding of academic challenges, it is crucial to collect these personal and contextual factors as part of the measurement process. This approach allows for a more nuanced interpretation of academic challenges, recognizing that they are influenced by both temporal progression through the semester and individual and contextual characteristic. A summary of proposed modifications based on scale-level examinations of the measure is presented in Table 19.

Table 19

Summary of Suggestions for Improving the Scale based on Scale-Level Analyses

Area of Improvement	Proposed Modification	Rationale
Granularity of the Scale	- Develop a shorter, less granular scale for measuring general academic challenges and more granular, independent scales for specific challenges. - add items to the specific scales with the potential to reflect unique aspects of each challenge domain that are likely to remain unaccounted for by the general factor	Enhances reliability and predictive utility of both general and specific factors
Ecological Validity	- Define general factor <i>a priori</i> by incorporating questions that directly measure it - Prioritize ecological validity by closely aligning the general factor with students' subjective experiences	Better represents students' real-world experiences. No specific factor is collapsed as a referent domain of g
Data Driven g Factor	Avoid including functionally non-equivalent domains; treat these as independent constructs	Ensures interpretable data-driven general factor
Theory Driven g Factor	Anchor the g factor within a particular domain only when there is solid theoretical support	The meaning of the g changes once the referent factor changes
Volitional Processes	Integrate volitional processes into a comprehensive Volitional Challenges factor	Potentially address functional non-equivalency by allowing challenges associated with managing time to form a discrete scale
Contextual & Individual Moderators	Collect data on potential temporal, contextual, and individual moderators	Validity of interpreting the scale's scores may depend on these contextual factors

Practical Recommendations for Using the Current Version of the Scale

Table 20 summarizes the recommendations for using the scale in practical and research contexts and for clustering individuals based on their academic challenges scores.

Table 20

A summary of Practical Recommendations for Using the Current Version of the Academic Challenges Scale

Research Question demands operationalization of...	Can researchers use SEM?	Recommendations
A Single Specific Challenge	Yes	Model it as a unidimensional construct
A Subset of specific challenges	Yes	Model them as unidimensional constructs that are allowed to covary (ideally using Correlated ESEM)
All specific challenges And/or General Academic Challenges	Yes	Consider modeling it as a S-1 bifactor ESEM with metacognitive challenges as a reference domain (note the low reliability of some subscales in the current version of the scale) *
Specific Challenges	No	Model as average subscale scores
Overall Academic Challenges	No	Model as an average total score
Shorter representation of the overall score needed	Yes/No	Consider using 8 items of the metacognitive challenges scale for a shorter but highly correlated representation of the overall score**
Clustering individuals	Yes***	Use LPA with specific factors as indicators (no need to include g, it is implied in the level-differentiation across profiles) OR Use norm-based or comparative cut-offs for clustering based on the total score
Clustering individuals	No***	Use norm-based or comparative cut-offs for clustering based on the total score

Note. * See the section titled “Prioritizing Ecological Validity in Defining the General Factor” for a discussion of limitations of this approach.

** See the section titled “Suggestions for Refining the Unidimensional General Academic Challenges Scale (AC-g) for a discussion of limitations of this approach.

*** Can researchers use Finite Mixture Modeling (e.g., LPA)?

Computing Manifest Scores for the Bifactor Structure of the Scale

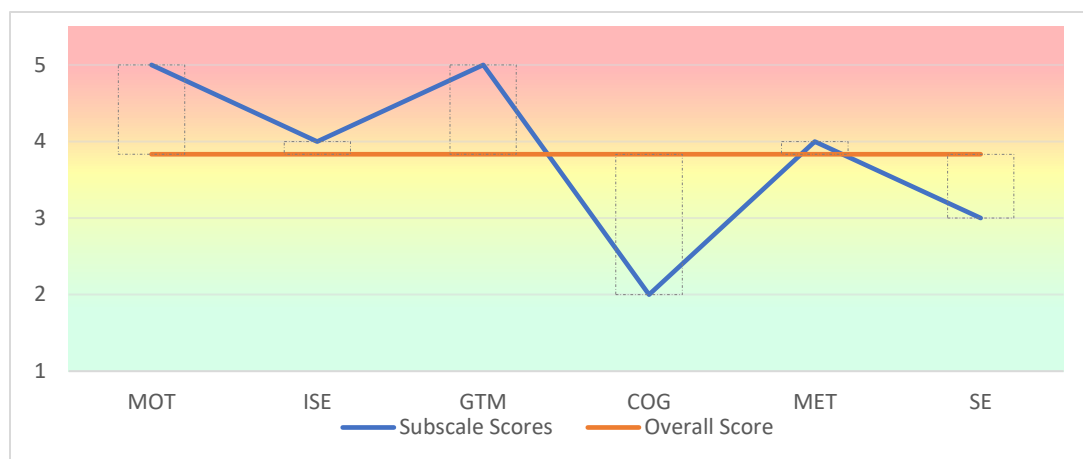
The practical application of the scale in settings where Structural Equation Modeling (SEM) is unavailable necessitates the development of simplified scoring methods using basic arithmetic. While Table 20 provides initial recommendations, further exploration of additional scoring opportunities is warranted.

In bifactor models, the interpretation and scoring of general and specific factors require careful consideration due to their unique psychometric properties and interrelationships. The general factor (g) typically represents the shared variance across all items. A practical approach to computing the g factor score could involve a weighted sum or average of all items, with weights reflecting the items' loadings on the g factor. This method ensures that items more representative of the common variance have a greater influence on the overall score.

Specific factors (S) scores, representing domain-specific variance after accounting for the g factor, can be calculated as average scores for each domain. However, their interpretation requires nuanced consideration. S factor scores should be viewed both independently and in relation to the g factor score. A higher S factor score relative to the g factor score indicates a positive deviation, suggesting that the specific challenge is more pronounced than expected based on the general level of challenges. In practical applications, it may be beneficial to interpret deviation scores for S factors, which represent the difference between an individual's score on a specific factor and their g factor score (as illustrated in Figure 23).

Figure 23

How to Interpret an Individual's Scores on a Specific factor in Relation to their Overall Score



Note. The green, yellow, and red background represent norm-based cut-offs for low, moderate, and high challenges. The dotted bars represent how much an individual's score on a specific factor deviates from their score on the g factor.

Computing non-refined composite scores in the S-1 bifactor model presents challenges due to the psychometric overlap between the referent factor and the general factor (g). While the scoring method described for the symmetrical bifactor model can be applied to the S-1 model, interpreting and using these scores should be done with caution. This is due to the inherent psychometric entanglement between the referent domain's score and the g factor score.

Additionally, given the very high association between the g factor and the referent factor, further investigation is needed to determine whether the referent factor alone can serve as a proxy for the general factor.

The complexity of disentangling general and specific factors in the S-1 model underscores the need for continued investigation into scoring methods that can accurately reflect the nuanced structure of academic challenges. It is important to acknowledge that the scoring methods suggested here does not fully separate specific and general challenges. Consequently, further research is needed to validate the interpretations based on this scoring method to ensure they provide meaningful and practically useful information.

Future Directions

Examining Students' Beliefs and Behaviours in the Context of Their Perceived Challenges

The *Academic Challenges* scale is intended for concurrent use with the other three measures in the Self-Regulated Learning Profiles and Self-Diagnostic tool (SRL-PSD): perceived importance of SRL, self-efficacy for SRL, and self-reported engagement in SRL practices. Together, these tools provide a detailed and relational profile of students' self-regulated learning, encompassing what they believe (importance of SRL and SRL self-efficacy), what they report they do (self-regulated learning practices), and the challenges they perceive (Study 1). This unified profile has the potential to facilitate the interpretation of students' learning practices and beliefs within the broader context of their challenges. Situating SRL practices and beliefs within the context of the challenges students face can potentially enhance understanding of how students intentionally respond to the challenges they perceive.

Person-centered analyses that explore how clusters of challenges, practices, and beliefs interrelate would also be valuable in providing a more comprehensive understanding of students' self-regulated learning experiences. However, before conducting such analyses, an extended psychometric analysis of the scales measuring SRL practices and beliefs should be conducted, similar to the evaluation conducted for the *Academic Challenges* scale. This analysis should aim to identify any potential limitations or areas for improvement in these scales, and modifications should be made accordingly. Then, person-centered analyses can be conducted to uncover distinct profiles of students based on the interrelationships between their self-reported challenges, practices, and beliefs related to self-regulated learning.

Additionally, as discussed, understanding potential interactions between contextual factors and SRL components in the context of academic challenges may help identify more

specific areas for intervention that are tailored to students' needs. Also, the idea of modeling academic challenges as a consequence of inadequate self-regulation, which subsequently triggers a more efficient phase of self-regulation, aligns with the cyclical nature of self-regulated learning (Hadwin et al., 2018; 2019; Järvelä et al., under review; Winne & Hadwin, 1998). To take such a dynamic nature into account, longitudinal designs such as experience sampling methods (see Wrzus, 2017) could be highly beneficial. Future research should explore how these aspects are associated with each other both cross-sectionally and longitudinally. This approach could involve pairing specific challenges with the practices students deploy and examining the efficacy of those practices for remediating challenges over time. Worsening challenges over time may be the result of (a) lack of awareness of practices for strategically responding to challenges, (b) sub-optimal selection of practices for the challenges at hand and (c) poor implementation and strategies or tactics. This line of inquiry represents an important future direction for research in understanding and addressing academic challenges.

Within Person and Cross-Context Reliability

The composite reliability and internal consistency of the academic challenges scale were examined in the studies included in this dissertation. However, to further enhance the utility of the scale, particularly in educational settings where academic challenges can fluctuate over time, evaluating within-person change, or test-retest reliability, is necessary.

Evaluating within-person change would enable researchers and educators to track how academic challenges evolve throughout different academic cycles, such as over a semester or an academic year. This can reveal if certain challenges are more prevalent at specific times, such as early or late in the semester, and whether these patterns are consistent across cohorts of students, disciplinary fields, or task specific situations. For instance, students who respond to the survey in

the second week may face social and emotional challenges, typical of early university life when adapting to new social settings is common. By the fourth week, students may perceive more cognitive and metacognitive challenges as academic demands increase. A repeated measurement approach would offer a dynamic view of how academic challenges change over time, potentially influenced by typical academic demands of each academic cycle and the accumulation of academic experiences, among others.

Repeated measurement can also shed light on the contextual factors that contribute to any possible changes in response patterns across measurement occasions. If this evaluation is conducted alongside the collection of data on potential contextual influences (e.g., changes in course load, personal stressors, type of tasks, etc.), it could provide a more comprehensive understanding of what factors most significantly impact students' academic experiences, offering valuable implications for both research and practice. For instance, it is important to determine how early in the semester students with certain challenge profiles likely to end up with poor performance and wellbeing can be identified by the scale, allowing for the provision of timely support.

A student's change in the challenges they perceive may depend on the external demands, but a student challenge profile can fluctuate if they change their standards, goals, effort, etc., even if the context remains unchanged. Consequently, collecting constructs that measure their expected performance and goals are also important to complete the picture. As mentioned earlier, incorporating questions that probe these temporal, contextual, and personal moderators and confounders can facilitate accumulating the knowledge gained from different studies in which challenges are investigated. Similarly, if the measure is intended for use in studies with a pretest–intervention–post-test design, considering these moderators in the model ensures that the

measure accurately reflects the dynamic nature of students' perceptions and can effectively gauge the impact of interventions. To facilitate the scale's usage in repeated assessments, developing parallel forms of the measure and establishing alternate-form reliability is helpful to mitigate practice effects on the measure. This approach ensures that repeated administration of the measure over time does not lead to inflated scores due to familiarity with the content.

Currently, the scale uses a two-week window to assess academic challenges. Assessing test-retest reliability can help determine whether this timeframe is sufficient for the proposed use of the scale. It is possible that a more prolonged measurement window could capture more persistent challenges, rather than transient ones that might not have a lasting impact on a student's overall academic performance and wellbeing.

Context Specific Use of the Scale

Considering an old debate in the self-regulated learning field regarding whether SRL is best measured as a domain-general, domain-specific, or situation-specific construct (Alexander et al., 2011), applying the *Academic Challenges* scale to specific domains could yield valuable implications. For example, by administering the scale in the context of individual courses, researchers and educators can identify the most prevalent challenges students encounter within each course, whether they pertain to cognition, motivation, or other areas. These targeted findings enable course-specific adjustments aimed at better addressing student needs and optimizing their learning experiences. Investigating variations in perceptions of challenges can also be done in a more granular task-level.

In sum, assessing stability of measure across potential personal, temporal, and contextual moderators is recommended as an important future direction.

Need for Person-Centered Study of Challenges

As discussed, the six *Academic Challenges* factors are highly interconnected, meaning that encountering challenges in one regulatory domain is likely to reflect challenges in other domains as well. For instance, procrastination in a student could arise from motivational issues such as low self-efficacy or not valuing the task sufficiently. Behaviorally, it might be worsened by poor study habits, such as studying in a socially distracting environment. Cognitively, it can manifest as an avoidance strategy for difficult tasks. From a metacognitive perspective, it may stem from inadequate skills in planning and setting attainable goals. From an emotion regulation viewpoint, it can be a coping response to high stress, or negative emotions linked to the task (see Steel, 2007 for a meta-analysis on procrastination). Consequently, interpretations and subsequent psychoeducational support should consider one's academic challenges as a relational profile rather than as a set of compartmentalized factors. This relational interconnectivity should also be considered when using the scales in research for hypothesis testing. For example, consider the following questions:

- What do *motivation* challenges represent as an independent factor?
- What does the covariation between *motivation* challenges and *initiating and sustaining engagement* (ISE) challenges indicate?
- What do *motivation* challenges represent after controlling for their covariation with *ISE* challenges?
- What are the meaningful ensembles of challenge factors that can be employed as stand-alone structural elements?

A person-centered analysis, such as latent profile analysis (see Collins & Lanza, 2009; Ding, 2018; Vermunt & Magidson, 2004), can significantly enhance our understanding of the

intraindividual interconnections among challenge areas and how these profiles differ in terms of factors in student success. Such an analysis could also aid in developing a categorical diagnostic approach for the scale based on different configurations of challenges in clusters, potentially facilitating better screening (see Keyes, 2002; Keyes et al., 2008 for an example of such an approach for a different scale).

Since the scale is intended for diagnostic purposes, future studies should explore the practical significance of the differences among the extracted profiles. Research should determine which levels of challenge or patterns within these profiles are critical and necessitate intervention. Additionally, these findings should be translated into original units of measurement for practical application. Understanding at what point and under what conditions support should be provided is essential for the effective use of the scale in educational settings.

The complementary person-centered study conducted as part of the Study 3 should be viewed as an exploratory investigation into the multidimensionality of the construct using a person-centered approach. Future studies should consider using clustering methods to identify specific groups of students who could benefit from tailored support strategies.

Examining the Ecological Validity of the Measure Using a Mixed Method Design

Psychometric studies can be complemented by qualitative interview studies within a mixed-methods design to corroborate the validation of a researcher's or diagnostician's interpretations of the scale. Interviews can be conducted after participants complete the questionnaire. During the interview, participants can be asked about their major academic challenges, and their responses can be compared with the survey results to assess alignment both in terms of the level and categorization of the challenges.

Additionally, students' responses to the interview questions can be contextualized within the items themselves. For instance, an item is first read aloud, and then participants are asked to explain why they selected a specific response. This approach helps understand what came to their mind when responding to the item. Participants can also provide examples to further contextualize their answers. A mix of open-ended and structured questions could be used to elicit detailed responses and gain deeper information into the participants' experiences and perspectives.

To further enrich this process, participants can be asked to explain the range of matters they considered when they selected a response option and why the response they selected (e.g., “often true”) is the best account. This can include discussing what “often” means to them and what situations or instances the participant acknowledged but decided not to consider in their overall response. By incorporating these qualitative insights, researchers can gain a deeper understanding of the nuances in participants' responses and potentially enhance the ecological validity of their interpretations of the measurement.

Triangulating with Other Data Modes

Methods of measuring SRL can be divided into two major categories. The first category involves understanding students' perceptions of their learning behaviours and experiences, which can be done in real time, such as with think-aloud protocols, or retrospectively via self-report questionnaires and interviews. The second category involves observational methods, such as tracking learners' actions in real time or traces of those actions during a study episode (e.g., using digital footprints). The primary assumption of self-report measures of self-regulated learning is that learners' self-regulatory behaviors are intentional, and they can accurately recall, aggregate, and report these behaviors. In contrast, the assumption for behavioral observation methods is that

specific observable behavioral patterns can be used to infer the intentional and purposive processes of self-regulated learning.

Self-report measures provide a means for assessing learners' perceptions of how they experience and manage their study behavior. However, these perceptions may not always align strongly with their actual behaviors (Winne & Jamieson-Noel, 2002). This potential discrepancy can arise due to various factors, including biases inherent to self-report measures, such as recall bias (Hassan, 2005) and social desirability bias (van de Mortel & Thea, 2008). Despite these potential pitfalls, self-report measures remain a valuable tool for gaining access to students' mental models of the types of challenges they encounter and how they think they typically regulate learning in response to these challenges. Regardless of how much these subjective perceptions deviate from behavioural data, appraisal of challenges and the subsequent strategy deployment is based on one's self-perceptions. In other words, a major part of self-regulation is managing one's subjective experience, the experiential quality of which is only accessible to the self. This is why one's perceptions and beliefs can be viewed as a causal element influencing behavior construction (Ajzen, 1991, 2020; Bandura, 1986; Dweck, 2006; Eccles & Wigfield, 2002; Weiner, 1985). A more holistic view of students' academic self-regulation emerges when the subjective experiences of learners are added to the investigation of other data modalities.

Other types of data, such as behavioral observations, can help verify whether students' self-reported perceptions are accurate, while self-reports can validate if inductive inferences from trace data truly reflect students' intents and perceptions. However, beyond such a cross-validation, discrepancies between self-reported experiences and other data forms can provide valuable insights into how students manage their learning. For example, we might find that some challenges are more salient in recollections of behaviours than in behavioural traces themselves.

Another hypothetical scenario could involve examining how differences between behavioral traces and self-reports relate to learners' metacognitive awareness and monitoring. One might speculate that learners with stronger monitoring skills would show fewer discrepancies.

In sum, in the evolving landscape of multi-modal study of learning, self-reports could serve as a vital modality, potentially offering unique insights into learners' subjective views and experiences that other data sources might not capture (see Järvelä & Bannert, 2019, for a special issue on understanding temporal and adaptive processes in learning regulation using multimodal data).

Simultaneous Modeling of Specific and General Factors In IRT

To address the objectives of Study 2 and minimize the overlap between item-level and scale-level analyses, Samejima's Graded Response Model (GRM; 1997, 2016), a unidimensional polytomous IRT model, was used in Study 2. Future research can employ a bifactor polytomous multidimensional item response model to better understand the general and specific aspects of the measure (see Immekus et al., 2019; Toland et al., 2017).

In this framework, statistical error arising from various sources is modeled, including imprecision in item parameter estimates, variability in estimating an individual's latent attribute on the general and specific factors, model misfit reflecting how well the bifactor structure represents the observed relationships in the data, and response inconsistency reflecting respondents' inconsistent answers across items from the same factor. The standard error of measurement in this context is typically calculated for each dimension (general and specific factors) and varies across the latent attribute continuum (Immekus et al., 2019; Toland et al., 2017).

Computerized Adaptive Scales

Investigating the sample invariant parameters of IRT is essential for the development of *computerized adaptive scales* (Dodd et al., 1995). These scales optimize the assessment process by tailoring the selection of items to each respondent's previous answers, rather than administering a fixed set of items to all participants. This customization should be driven by an algorithm that is both theoretically grounded and refined through empirical evidence (Hol et al., 2007).

The potential to develop a *computerized adaptive scale* for assessing academic challenges, particularly for creating personalized, self-diagnostic profiles, is promising. A long-standing debate in the field of self-regulated learning measurement concerns the appropriate level of measurement granularity (Winne, 2010b). Also, the need for different granularity levels for measuring specific and general factors was discussed earlier. Adaptive scales may offer potential solutions by allowing for a tailored balance between the breadth and depth of the assessment. For example, if a person does not perceive motivation challenges, instead of asking detailed questions about this area, it might be more effective to concentrate on specific details in other challenge areas. This approach has the potential to create detailed personalized profiles of challenges while reducing response fatigue. Future research should also examine potential theoretical frameworks and algorithms that can operationalize these adaptive scales. Table 21 outlines the key recommendations for future research directions.

Table 21*Summary of Future Research Directions*

Future Directions	Details
Examining Students' Beliefs and Behaviors in the Context of Their Perceived Challenges	Integrate the Academic Challenges scale with other SRL-PSD measures for a comprehensive profile of self-regulated learning.
Within Person and Cross-Context Reliability	Investigate stability of the measure across potential personal, temporal, and contextual confounders or
Context-Specific Use of the Scale	Apply the scale to specific domains, courses, or tasks to identify prevalent challenges.
Need for Person-Centered Study of Challenges	Use person-centered analyses like latent profile analysis to uncover distinct profiles of challenges, practices, and beliefs. Explore practical significance of profile differences for targeted interventions.
Examining the Ecological Validity of the Measure Using a Mixed-Method Design	Complement psychometric studies with qualitative interviews to corroborate the ecological validity of the interpretation based on the scale
Triangulating with Other Data Modes	Use self-report measures and behavioral observations to cross-validate findings. Explore differences in perceived and actual behaviors. <u>Examine discrepancies to gain insights into self-regulation.</u>
Simultaneous Modeling of Specific and General Factors in IRT	Employ bifactor polytomous multidimensional IRT models to gain insights into general and specific aspects of the measure.
Computerized Adaptive Scales	Develop computerized adaptive scales to tailor item selection based on responses. Balance assessment breadth and depth to reduce response fatigue. Explore theoretical frameworks and algorithms for operationalizing adaptive scales.

Limitations

Despite the insights gained from this study, several limitations should be acknowledged. Addressing these limitations can guide future research to build upon and enhance the findings presented here.

COVID Context. The data for this study were gathered during the COVID-19 pandemic, specifically in the Fall of 2021 and Winter of 2022. During this period, all participating students engaged in university courses via online, off-campus education. For the majority of both students and instructors, online instruction was a relatively new mode of learning and teaching. The students were geographically dispersed across Canada and internationally, with most studying from their homes. The extent of COVID-19 restrictions on socializing varied both within Canada

and around the world. Furthermore, many extracurricular activities were either canceled or faced substantial limitations. These unique circumstances should be taken into consideration when interpreting the study's results. Future replication studies to further validate these findings under different conditions would be beneficial.

Samples. While 4 student samples were used for studies included in this dissertation, the same dataset was used for Study 2 and Study 3. This decision should not be interpreted as data overuse, as these two studies are complementary in nature. Study 2 focused on an item-level examination using graded response IRT to assess measurement precision of individual items. In contrast, Study 3 explored the multidimensionality of the construct. These distinct focuses—item-level precision across different ranges of academic challenges and factorial structure—provided a complementary understanding of the scale by maximizing insights through different analytical lenses.

Furthermore, full measurement invariance was observed between the two constituting cohorts of the sample (Fall intake N=1246, Winter intake N=1427). Findings from the measurement invariance supported *configural*, *metric*, *scalar*, *strict*, latent variance-covariance, and latent means invariance across these two cohorts and indicated that even if the analyses were conducted with different cohorts (e.g., one cohort for Study 2 and another for Study 3), the findings would have been consistent. Using the same sample for the two studies allowed for examining the methodological synergies of the analytical methods employed.

Additionally, Study 1, which used two separate samples, yielded comparable results. This consistency across different samples lends credence to the reliability of findings and suggests that the findings may be applicable to analogous populations.

However, researchers must be cautious to avoid overfitting models to a specific dataset, a risk that increases with repeated data use. Overfitting can make models too tailored to one set of data, potentially reducing their relevance to other datasets or real-world scenarios.

Though robust standard error was used for all analyses, multiple analyses, such as examining the associations of the academic challenges with criterion variables across nested measurement models, may have increased the probability of Type I error (the chances of finding significant results purely by chance). These analyses should be regarded as descriptive information about the utility of different analytic approaches, with generalizability inferences warranting replications. Future studies should use new samples with diverse characteristics to test replicability of the findings.

Grade Point Average. GPA from the participating universities was converted into a 5-point scale to match the less detailed GPA scales of two participating universities. The numerical values were converted to the letter grades based on the site-specific guidelines provided by each institution to ensure the meaning of GPA remains consistent even though the granularity changes. However, this conversion could have potentially distorted the relative performance levels of students, particularly if the original scales differ significantly in granularity.

Furthermore, the distribution of GPA data across these public universities indicated that the sample predominantly consisted of high-performing students (range=1 to 5; median=4). This skewness towards high performers limits the generalizability of the findings to populations with different academic performance characteristics. Future studies should employ a stratified sampling or employ weighting methods to ensure a more representative sample.

Moreover, due to having a separate consent procedure for collecting GPA data, the extent of missing GPA data was noticeable (33% missing) and differed across sites, due to variations in

consent processes and institutional regulations. This level of missing data can introduce bias into the result, especially if the missingness is systematic (e.g., lower-performing students being less likely to consent to their GPA data being collected). Calculating semester GPA was also challenging, given the different frequencies of year-long courses across sites. In these courses, part of the GPA awarded in the second semester is, in fact, attributable to the work completed in the previous semester.

Despite all these limitations, when the findings from this sample was compared to Study 1, in which the measure was administered towards the end of an academic semester to 476 students from a single university, the findings aligned, although the effect sizes in Study 1 were marginally larger.

Criterion Constructs. In the studies included in this dissertation, the criterion utility of the *Academic Challenges* scale was examined against students' expected academic performance, self-efficacy for performance, and actual academic performance (end-of-semester GPA).

The calibration between students' expectations and their actual performance is another crucial variable that was not examined in these studies. Understanding this alignment can shed light on students' metacognitive monitoring and self-evaluation skills. Future research could also collect data on students' intended performance ranges to interpret their behavior and experiences in the context of their personal standards rather than normative criteria. This approach would offer a more individualized understanding of student performance.

Future studies could also consider incorporating a broader range of criterion constructs at various levels of proximity that potentially mediate the link between academic challenges and academic outcomes, student success, and well-being. This approach could provide a more comprehensive understanding of the factors associated with academic challenges by including

aspects of both well-being and performance. Additionally, it may help to develop a model of potential correlates at varying levels of proximity that may serve as mediators between challenges and desired outcomes.

Measurement Invariance. Studies 2 and 3 employed differential item functioning (DIF) and multigroup SEM Measurement Invariance (MI) testing, respectively, to assess the stability of psychometric parameters across various demographic groups. These groups included five data collection sites, two cohorts, sex at birth, and the student's year of study. The analyses established insensitivity of measurement parameters to those variables, indicating that these demographic variables did not introduce significant variance in the measurement parameters.

The selection of these demographic characteristics was driven by practical considerations. Cohorts and sites were examined to validate the combination of subsamples into an integrated sample. Year of study was investigated due to potential variations in academic challenges between first year and other students. Sex at birth was analyzed to investigate whether the measure was unbiased across assigned sex. Although these choices were justified, they can invite questions about why other variables (e.g., socioeconomic status) were not examined. It is important to note that testing MI across these particular variables does not mean that other untested variables are unimportant. These demographic variables were selected based on practical constraints and methodological considerations in the larger study providing data for this dissertation. These choices were influenced by data availability, the complexity of the measurement models, and the characteristics of the demographic variables themselves. Some variables were multicategorical, had disproportionate sample sizes across categories, or showed limited occurrence in certain levels, making MI testing challenging within the scope of our study.

Furthermore, although establishing invariance across demographic variables is important and can have important social implications, foci for investigating measurement invariance can and perhaps should extend beyond these categories. The selection of contextual or individual variables for MI testing should be grounded in theory, encompassing potential proximate causes that might bias measurement. This approach is important for a comprehensive scientific understanding of the sample.

From a measurement perspective, identifying potential confounders or moderators that bias psychometric parameters is as crucial as establishing measurement invariance. Each sample can be viewed as comprising numerous subsamples and determining the appropriate level of granularity for sample description is essential. The definition of a representative sample and the suitable level of granularity for potential moderators and confounders should be theoretically informed. Consequently, MI cannot be ensured based solely on universal demographic characteristics but must also consider the nature of the phenomena being measured and other theory-delineated factors.

An illustrative example involves a study of academic challenges across sex at birth. Initial MI testing across the genetically narrow but socioculturally broad categories of male and female might suggest equivalence. However, a more granular examination considering intersections with factors like academic discipline (e.g., male in STEM, female in STEM) could reveal significant non-invariance.

This scenario underscores the importance of considering intersectionality and complex groupings in MI testing, rather than relying solely on broad demographic categories. Several advanced statistical methods can be used for examining MI across intersections of multiple variables. These include Multiple-Group Confirmatory Factor Analysis (MG-CFA), Moderated

Nonlinear Factor Analysis (MNLFA), the Alignment Method, Multilevel Factor Analysis, and Multiple Indicators Multiple Causes (MIMIC) models. These methods provide a robust toolkit for investigating MI in complex, intersectional contexts.

Also, when invariance is not achieved, researchers should explore adjusting the level of granularity based on theory to identify more proximate causes of bias. For instance, initial MI testing across age groups might fail, but regrouping based on levels of competing life responsibilities could achieve invariance. This example demonstrates how a demographic variable like age can mask a more proximal factor causing measurement non-invariance.

In the studies included in this dissertation, measurement invariance was established across the tested demographic variables. However, future research should consider incorporating more theory-driven and proximate variables that may causally bias measurement.

Conclusion

Students often face diverse challenges throughout their educational journey that can significantly impact their learning process. While extensive research has explored self-regulatory perceptions and behaviors that promote student success, there remains a notable gap in understanding the specific challenges students encounter during their studies. Viewing these challenges through a self-regulatory lens is crucial, as it not only identifies barriers but also reveals key opportunities for developing more refined and effective learning strategies. These challenges, spanning motivational, behavioral, volitional, cognitive, metacognitive, social, and emotional dimensions, are essential to identify and understand from the students' perspective.

Learners' perceptions of challenge experiences, among other factors, shape their choices in managing their studies. Thus, managing one's subjective experience constitutes a crucial

aspect of self-regulation. Given the subjective nature of challenges, self-report measures offer a valuable tool for evaluating learners' perceptions of their academic challenges experiences.

In the studies included in this dissertation, both item-level analysis, such as polytomous Graded Response Model Item Response Theory (GRM IRT), and scale-level analyses of multidimensional structure, including confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and bifactor modeling were employed. A person-centered study was conducted as well to further examine the multidimensional structure and clustering based on the measure. These approaches were selected to provide complementary insights about the *Academic Challenges* scale as a potential self-assessment and diagnostic tool, using four student samples. Results demonstrated adequate psychometric properties of the *Academic Challenges* scale across multiple psychometric criteria, including reliability, and factor structure. However, adhering to a view of measurement as a continuous improvement process, recommendations for potential item-level and scale-level refinement in future iterations of the measure were provided.

To conclude, referring back to the opening quote of this chapter, while psychometric properties are what can be counted in the measurement process, ensuring that what's been counted counts requires consideration of other layers of measurement quality that cannot be readily counted but do count. Testing the suggested modifications against empirical evidence should be complemented with further theoretical examinations to ensure that what is statistically significant is also practically significant.

References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Ajzen, I. (2020). The theory of planned behavior: Frequently asked questions. *Human Behavior and Emerging Technologies*, 2(4), 314-324. <https://doi.org/10.1002/hbe2.195>
- Alexander, P. A., Dinsmore, D. L., Parkinson, M. M., & Winters, F. I. (2011). Self-regulated learning in academic domains: University of Maryland. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 407-421). Routledge. <https://doi.org/10.4324/9780203839010.CH25>
- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 397-438. <https://doi.org/10.1080/10705510903008204>
- Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 495-508. <https://doi.org/10.1080/10705511.2014.919210>
- Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management*, 41(6), 1561-1577. <https://doi.org/10.1177/0149206315591075>
- Aybek, E. C., & Demirtasli, R. N. (2017). Computerized Adaptive Test (CAT) Applications and Item Response Theory Models for Polytomous Items. *International Journal of Research in Education and Science*, 3(2), 475-487. <https://doi.org/10.21890/IJRES.327907>
- Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, 96, 207-210. <https://doi.org/10.1016/j.chb.2019.03.025>
- Bahena-Olivares, L. M., Rostampour, R., & Hadwin, A. F. (2024). Modelling the unique contributions of task understanding to academic challenges and performance [Manuscript submitted for publication]. University of Victoria.
- Baker, F. B. (2001). The basics of item response theory. For full text: <http://ericae.net/irt/baker>
- Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R* (Vol. 969). Springer. <https://doi.org/10.1007/978-3-319-54205-8>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

- Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. W.H. Freeman and Company.
- Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52(1), 1-26. <https://doi.org/10.1146/ANNUREV.PSYCH.52.1.1>
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497-529. <https://doi.org/10.1037/0033-2909.117.3.497>
- Beauducel, A. (2011). Indeterminacy of factor score estimates in slightly misspecified confirmatory factor models. *Journal of Modern Applied Statistical Methods*, 10(2), 583-598. <https://doi.org/10.56801/10.56801/v10.i.566>
- Bhattacharjee, A. (2012). *Social science research: Principles, methods, and practices*. University of South Florida.
- Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, 7(2), 161-186. <https://doi.org/10.1016/S0959-4752%2896%2900015-1>
- Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, 16(3), 265-284. <https://doi.org/10.1037/a0024448>
- Bollen, K. A., & Lennox, R. D. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305-314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science*, 5(1), 184-186. <https://doi.org/10.1177/2167702616657069>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.
- Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, 80(4), 796-846. <https://doi.org/10.1111/j.1467-6494.2011.00749.x>
- Bureau, J. S., Guay, F., Plamondon, A., Ratelle, C. F., Howard, J. L., & Gilbert, W. (2023). Empirical testing of an alternative modeling of the self-determination continuum. *Motivation and Emotion*, 47(1), 46-60. <https://doi.org/10.1007/s11031-022-09976-9>
- Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020). Application of the bifactor S–1 model to multisource ratings of ADHD/ODD symptoms: An appropriate bifactor model for symptom ratings. *Journal of Abnormal Child Psychology*, 48, 881-894. <https://doi.org/10.1007/s10802-019-00608-4>

- Cambridge-Williams, T., Winsler, A., Kitsantas, A., & Bernard, E. (2013). University 100 Orientation Courses and Living-Learning Communities Boost Academic Retention and Graduation via Enhanced Self-Efficacy and Self-Regulated Learning. *Journal of College Student Retention*, 15(2), 243-268. <https://doi.org/10.2190/CS.15.2.f>
- Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved Differential Test Functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, 76(1), 114-140. <https://doi.org/10.1177/0013164415584576>
- Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E. L., Van der Kaap-Deeder, J., Duriez, B., Lens, W., Matos, L., & Mouratidis, A. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*, 39, 216-236. <https://doi.org/10.1007/S11031-014-9450-1>
- Choi, J. N., & Moran, S. V. (2009). Why not procrastinate? Development and validation of a new active procrastination scale. *The Journal of Social Psychology*, 149(2), 195-212. <https://doi.org/10.3200/SOCP.149.2.195-212>
- Chun Chu, A. H., & Choi, J. N. (2005). Rethinking procrastination: Positive effects of "active" procrastination behavior on attitudes and performance. *The Journal of Social Psychology*, 145(3), 245-264. <https://doi.org/10.3200/SOCP.145.3.245-264>
- Church, A. T., Katigbak, M. S., Locke, K. D., Zhang, H., Shen, J., de Jesús Vargas-Flores, J., Ibáñez-Reyes, J., Tanaka-Matsumi, J., Curtis, G. J., & Cabrera, H. F. (2013). Need satisfaction and well-being: Testing self-determination theory in eight cultures. *Journal of Cross-Cultural Psychology*, 44(4), 507-534. <https://doi.org/10.1177/0022022112466590>
- Cleary, T. J., & Zimmerman, B. (2004). Self-Regulation Empowerment Program: A School-Based Program to Enhance Self-Regulated and Self-Motivated Cycles of Student Learning. *Psychology in the Schools*, 41(5), 537-550. <https://doi.org/10.1002/PITS.10177>
- Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, 12(4), 381-398. <https://doi.org/10.1037/1082-989X.12.4.381>
- Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons. <https://doi.org/10.1002/9780470567333>
- Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research*, 61(12), 1250-1262. <https://doi.org/10.1016/J.JBUSRES.2008.01.013>

- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119(2), 166.e7-166.e16. <https://doi.org/10.1016/J.AMJMED.2005.10.036>
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.
- Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, 10(1), 7. <https://doi.org/10.7275/JYJ1-4868>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. <https://doi.org/10.1037/H0040957>
- Davis, S. K., & Hadwin, A. F. (2021). Exploring differences in psychological well-being and self-regulated learning in university student success. *Frontline Learning Research*, 9(1), 30-52. <https://doi.org/10.14786/FLR.V9I1.581>
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
- de Bruijn-Smolters, M., Timmers, C. F., Gawke, J. C. L., Schoonman, W., & Born, M. P. (2016). Effective self-regulatory processes in higher education: Research findings and future directions. A systematic review. *Studies in Higher Education*, 41(1), 139-158. <https://doi.org/10.1080/03075079.2014.915302>
- Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, 19(2), 109-134. <https://doi.org/10.1016/0092-6566%2885%2990023-6>
- Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 416-436). Sage Publications Ltd. <https://doi.org/10.4135/9781446249215.n21>
- DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.
- Diamantopoulos, A., & Winklhofer, H. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, 38(2), 269-277. <https://doi.org/10.1509/jmkr.38.2.269.18845>
- Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61(12), 1203-1218. <https://doi.org/10.1016/J.JBUSRES.2008.01.009>

- Ding, C. S. (2018). *Fundamentals of applied multidimensional scaling for educational and psychological research*. Springer. https://doi.org/10.1007/978-3-319-78172-3_10
- DiStefano, C., Zhu, M., & Mindrila, D. (2019). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, 14(1), 20. <https://doi.org/10.7275/DA8T-4G52>
- Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19(1), 5-22. <https://doi.org/10.1177/014662169501900103>
- Dueber, D. M., & Toland, M. D. (2023). A bifactor approach to subscore assessment. *Psychological Methods*, 28(1), 222. <https://doi.org/10.1037/met0000459>
- Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In *Psychology of learning and motivation* (Vol. 54, pp. 103-140). Academic Press. <https://doi.org/10.1016/B978-0-12-385527-5.00004-8>
- Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.
- Dweck, C. (2015). Carol Dweck revisits the growth mindset. *Education Week*, 35(5), 20-24.
- Dweck, C. (2016). What having a "growth mindset" actually means. *Harvard Business Review*, 13(2), 2-5.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1), 109-132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, 101859. <https://doi.org/10.1016/j.cedpsych.2020.101859>
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5(2), 155-174. <https://doi.org/10.1037/1082-989X.5.2.155>
- Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, 23(1), 138. <https://doi.org/10.1037/met0000121>
- Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, 1(1), 3-14. <https://doi.org/10.1016/J.EDUREV.2005.11.001>

- Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, 13(4), 277-287.
<https://doi.org/10.1027/1016-9040.13.4.277>
- Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, 46(1), 6-25.
<https://doi.org/10.1080/00461520.2011.538645>
- Eid, M. (2020). Multi-Faceted Constructs in Abnormal Psychology: Implications of the Bifactor S - 1 Model for Individual Clinical Assessment. *Journal of Abnormal Child Psychology*, 48, 895-900. <https://doi.org/10.1007/s10802-020-00624-9>
- Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous Results in G-Factor Models: Explanations and Alternatives. *Psychological Methods*, 22(3), 541-562.
<https://doi.org/10.1037/met0000083>
- Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence*, 6(3), 42. <https://doi.org/10.3390/jintelligence6030042>
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Enders, C. K. (2022). *Applied missing data analysis*. Guilford Publications.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, 45(3), 225-245. <https://doi.org/10.1111/J.1745-3984.2008.00062.X>
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research. *Social Psychological and Personality Science*, 8(4), 370-378.
<https://doi.org/10.1177/1948550617693063>
- Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The Nature of Intelligence* (pp. 231-235). Lawrence Erlbaum Associates.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906-911.
<https://doi.org/10.1037/0003-066X.34.10.906>

- Flora, D. B. (2020). Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484-501.
<https://doi.org/10.1177/2515245920951747>
- Furr, R. M. (2021). *Psychometrics: An introduction*. SAGE Publications.
- Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). Sage Publications.
- Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science*, 50(1), 21-43.
- Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, 55, 57-68.
<https://doi.org/10.1016/J.INTELL.2016.01.006>
- Gillet, N., Morin, A. J., Huart, I., Colombat, P., & Fouquereau, E. (2019). The Forest and the Trees: Investigating the Globality and Specificity of Employees' Basic Need Satisfaction at Work. *Journal of Personality Assessment*, 102(5), 702-713.
<https://doi.org/10.1080/00223891.2019.1591426>
- Giordano, C., & Waller, N. G. (2020). Recovering bifactor models: A comparison of seven methods. *Psychological Methods*, 25(2), 143-156. <https://doi.org/10.1037/met0000227>
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1), 26-34. <https://doi.org/10.1037/0003-066X.48.1.26>
- Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. Routledge.
- Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, 77(3), 334-372. <https://doi.org/10.3102/003465430303953>
- Greene, J. A., Robertson, J., & Costa, L. J. C. (2011). Assessing Self-Regulated Learning Using Think-Aloud Methods: University of North Carolina at Chapel Hill. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 327-342). Routledge. <https://doi.org/10.4324/9780203839010.CH20>
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430-450.
- Hadwin, A. F., Davis, S. K., Bakhtiar, A., & Winne, P. H. (2019). Academic challenges as opportunities to learn to self-regulate learning. In H. Askell-Williams & J. Orrell (Eds.), *Problem solving for teaching and learning* (pp. 34-47). Routledge.
<https://doi.org/10.4324/9780429400902-4>

- Hadwin, A. F., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 83-106). Routledge. <https://doi.org/10.4324/9781315697048.CH6>
- Hadwin, A. F., Winne, P. H., Stockley, D. B., Nesbit, J. C., & Woszczyna, C. (2001). Context moderates students' self-reports about how they study. *Journal of Educational Psychology*, 93(3), 477-487. <https://doi.org/10.1037//0022-0663.93.3.477>
- Hair, J. F., Hult, G. T., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017). Mirror, mirror on the wall: a comparative evaluation of composite-based structural equation modeling methods. *Journal of the Academy of Marketing Science*, 45(5), 616-632. <https://doi.org/10.1007/s11747-017-0517-x>
- Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 195-216). Scientific Software International.
- Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., & Kirby, J. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, 95, 103208. <https://doi.org/10.1016/j.jbi.2019.103208>
- Hassan, E. (2005). Recall Bias can be a Threat to Retrospective and Prospective Research Designs. *The Internet Journal of Epidemiology*, 3(2), 4.
- Hayes, A. F., & Rockwood, N. J. (2020). Conditional Process Analysis: Concepts, Computation, and Advances in the Modeling of the Contingencies of Mechanisms. *American Behavioral Scientist*, 64(1), 19-54. <https://doi.org/10.1177/0002764219859633>
- Haywood, D., Baughman, F. D., Mullan, B. A., & Heslop, K. R. (2021). Going "Up" to Move Forward: S-1 Bifactor Models and the Study of Neurocognitive Abilities in Psychopathology. *International Journal of Environmental Research and Public Health*, 18(14), 7413. <https://doi.org/10.3390/ijerph18147413>
- Heinrich, M., Geiser, C., Zagorscak, P., Burns, G. L., Bohn, J., Becker, S. P., Eid, M., & Knaevelsrud, C. (2023). On the meaning of the "P factor" in symmetrical bifactor models of psychopathology: Recommendations for future research from the bifactor-(S- 1) perspective. *Assessment*, 30(3), 487-507. <https://doi.org/10.1177/10731911211060298>

- Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck depression inventory–II. *Assessment*, 27(7), 1429-1447. <https://doi.org/10.1177/1073191118803738>
- Hermida, R. (2015). The problem of allowing correlated errors in structural equation modeling: concerns and considerations. *Computational Methods in Social Sciences*, 3(1), 5-17.
- Higgins, E. T. (1998). Promotion and Prevention: Regulatory Focus as A Motivational Principle. *Advances in Experimental Social Psychology*, 30, 1-46. <https://doi.org/10.1016/S0065-2601%2808%2960381-0>
- Higgins, E. T. (2014). Beyond pleasure and pain: Value from engagement. In J. Y. Shah & W. L. Gardner (Eds.), *Handbook of motivation science* (pp. 21-34). Guilford Press.
- Higgins, E. T., Kruglanski, A. W., & Pierro, A. (2003). Regulatory Mode: Locomotion and Assessment as Distinct Orientations. *Advances in Experimental Social Psychology*, 35, 293-344. [https://doi.org/10.1016/S0065-2601\(03\)01005-0](https://doi.org/10.1016/S0065-2601(03)01005-0)
- Hol, A. M., Vorst, H. C., & Mellenbergh, G. J. (2007). Computerized Adaptive Testing for Polytomous Motivation Items: Administration Mode Effects and a Comparison With Short Forms. *Applied Psychological Measurement*, 31(5), 412-429. <https://doi.org/10.1177/0146621606297314>
- Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, 4, 45. <https://doi.org/10.3389/feduc.2019.00045>
- Jansen, R. S., Van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (2019). Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. *Educational Research Review*, 28, 100292. <https://doi.org/10.1016/j.edurev.2019.100292>
- Jaramillo, F., & Spector, P. E. (2004). The effect of action orientation on the academic performance of undergraduate marketing majors. *Journal of Marketing Education*, 26(3), 250-260. <https://doi.org/10.1177/0273475304268780>
- Järvelä, S., & Bannert, M. (2019). Temporal and adaptive processes of regulated learning - What can multimodal data tell? *Learning and Instruction*, 72, 101268. <https://doi.org/10.1016/j.learninstruc.2019.101268>
- Järvelä, S., Malmberg, J., Haataja, E., Sobocinski, M., & Kirschner, P. A. (2021). What multimodal data can tell us about the students' regulation of their learning process? *Learning and Instruction*, 72, 101203. <https://doi.org/10.1016/j.learninstruc.2019.04.004>

- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. *Journal of Consumer Research*, 30(2), 199-218. <https://doi.org/10.1086/376806>
- Jobst, L. J., Bader, M., & Moshagen, M. (2023). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods*, 28(1), 207. <https://doi.org/10.1037/met0000423>
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American Statistical Association*, 70(351), 631-639. <https://doi.org/10.1080/01621459.1975.10482485>
- Jostmann, N. B., & Koole, S. L. (2010). Dealing with high demands: The role of action versus state orientation. In R. H. Hoyle (Ed.), *Handbook of personality and self-regulation* (pp. 332-352). Wiley-Blackwell. <https://doi.org/10.1002/9781444318111.CH15>
- Kalliath, T., & Brough, P. (2008). Work–life balance: A review of the meaning of the balance construct. *Journal of Management & Organization*, 14(3), 323-327. <https://doi.org/10.1017/S1833367200003308>
- Kang, H. A., Su, Y. H., & Chang, H. H. (2018). A note on monotonicity of item response functions for ordered polytomous item response theory models. *British Journal of Mathematical and Statistical Psychology*, 71(3), 523-535. <https://doi.org/10.1111/bmsp.12131>
- Kankaraš, M., Vermunt, J. K., & Moors, G. B. (2011). Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches. *Sociological Methods & Research*, 40(2), 279-310. <https://doi.org/10.1177/0049124111405301>
- Karabenick, S. A., & Berger, J.-L. (2013). Help seeking as a self-regulated learning strategy. In H. Bembenutty, T. J. Cleary, & A. Kitsantas (Eds.), *Applications of self-regulated learning across diverse disciplines: A tribute to Barry J. Zimmerman* (pp. 237-261). IAP Information Age Publishing.
- Karabenick, S. A., & Gonida, E. N. (2017). Academic help seeking as a self-regulated learning strategy: Current issues, future directions. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 421-433). Routledge.
- Karimi, L., & Meyer, D. (2014). Structural equation modeling in psychology: the history, development and current challenges. *International Journal of Psychological Studies*, 6(4), 123-133. <https://doi.org/10.5539/IJPS.V6N4P123>

- Kazén, M., Kaschel, R., & Kuhl, J. (2008). Individual differences in intention initiation under demanding conditions: Interactive effects of state vs. action orientation and enactment difficulty. *Journal of Research in Personality*, 42(3), 693-715. <https://doi.org/10.1016/J.JRP.2007.09.005>
- Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, 43(2), 207-222. <https://doi.org/10.2307/3090197>
- Keyes, C. L. M. (2005). Mental illness and/or mental health? Investigating the axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, 73(3), 539-548. <https://doi.org/10.1037/0022-006X.73.3.539>
- Keyes, C. L. M., & Annas, J. (2009). Feeling good and functioning well: Distinctive concepts in ancient philosophy and contemporary science. *The Journal of Positive Psychology*, 4(3), 197-201. <https://doi.org/10.1080/17439760902844228>
- Keyes, C. L. M., Wissing, M., Potgieter, J. P., Temane, M., Kruger, A., & van Rooy, S. (2008). Evaluation of the mental health continuum-short form (MHC-SF) in Setswana-speaking South Africans. *Clinical Psychology & Psychotherapy*, 15(3), 181-192. <https://doi.org/10.1002/cpp.572>
- Kim, Y., Brady, A. C., & Wolters, C. A. (2020). College students' regulation of cognition, motivation, behavior, and context: Distinct or overlapping processes? *Learning and Individual Differences*, 80, 101872. <https://doi.org/10.1016/j.lindif.2020.101872>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Knoesen, R., & Naudé, L. (2018). Experiences of flourishing and languishing during the first year at university. *Journal of Mental Health*, 27(3), 269-278. <https://doi.org/10.1080/09638237.2017.1370635>
- Koivuniemi, M., Panadero, E., Malmberg, J., & Järvelä, S. (2017). Higher education students' learning challenges and regulatory skills in different learning situations. *Infancia y Aprendizaje*, 40(1), 19-55. <https://doi.org/10.1080/02103702.2016.1272874>
- Kuhl, J., & Beckmann, J. (1994). *Volition and personality: Action versus state orientation*. Hogrefe.
- Kuhl, J., Quirin, M., & Koole, S. L. (2020). The functional architecture of human motivation: Personality systems interactions theory. *Advances in Motivation Science*, 7, 1-43. <https://doi.org/10.1016/bs.adms.2020.06.001>
- Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods*, 13(3), 435-455. <https://doi.org/10.1177/1094428109352528>

- Lance, C. E., & Vandenberg, R. J. (2010). On the practice of allowing correlated residuals among indicators in structural equation models. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 213-236). Routledge. <https://doi.org/10.4324/9780203867266-16>
- Landis, S. K., Sherman, M. F., Piedmont, R. L., Kirkhart, M. W., Rapp, E. M., & Bike, D. H. (2009). The relation between elevation and self-reported prosocial behavior: Incremental validity over the five-factor model of personality. *The Journal of Positive Psychology*, 4(1), 71-84. <https://doi.org/10.1080/17439760802399208>
- LimeSurvey GmbH. (n.d.). *LimeSurvey: An Open Source survey tool*. <http://www.limesurvey.org>
- Magidson, J., Vermunt, J. K., & Madura, J. P. (2020). Latent class analysis. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), *SAGE Research Methods Foundations*. SAGE Publications Limited. <http://dx.doi.org/10.4135/9781526421036>
- Mai, Y., Zhang, Z. J., & Wen, Z. (2018). Comparing Exploratory Structural Equation Modeling and Existing Approaches for Multiple Regression with Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(5), 737-749. <https://doi.org/10.1080/10705511.2018.1444993>
- Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275-340). Lawrence Erlbaum Associates Publishers.
- Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85-110. <https://doi.org/10.1146/annurev-clinpsy-032813-153700>
- Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 439-476. <https://doi.org/10.1080/10705510903008220>
- Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, 49(6), 1194-1218. <https://doi.org/10.1037/a0026913>
- Martin, J. (2004a). Self-regulated learning, social cognitive theory, and agency. *Educational Psychologist*, 39(2), 135-145. https://doi.org/10.1207/s15326985ep3902_4
- Martin, J. (2004b). The educational inadequacy of conceptions of self in educational psychology. *Interchange*, 35, 185-208. <https://doi.org/10.1007/BF02698849>

- Martin, J. (2007). The selves of educational psychology: Conceptions, contexts, and critical considerations. *Educational Psychologist*, 42(2), 79-89.
<https://doi.org/10.1080/00461520701263244>
- Martin, J., & McLellan, A. (2008). The Educational Psychology of Self-Regulation: A Conceptual and Critical Analysis. *Studies in Philosophy and Education*, 27, 433-448.
<https://doi.org/10.1007/S11217-007-9060-4>
- Martin, J., & McLellan, A.-M. (2013). *The education of selves: How psychology transformed students*. Oxford University Press.
<https://doi.org/10.1093/acprof%3Aoso%2F9780199913671.001.0001>
- McCombs, B. L., & Marzano, R. J. (1990). Putting the Self in Self-Regulated Learning: The Self as Agent in Integrating Will and Skill. *Educational Psychologist*, 25(1), 51-69.
https://doi.org/10.1207/S15326985EP2501_5
- McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, 52(5), 509-516. <https://doi.org/10.1037/0003-066X.52.5.509>
- McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.
<https://doi.org/10.2307/2669496>
- McNeish, D. M., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52, 2287-2305. <https://doi.org/10.3758/s13428-020-01398-0>
- Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, 95(4), 728-743.
<https://doi.org/10.1037/a0018966>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543. <https://doi.org/10.1007/BF02294825>
- Messick, S. J. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, 18(2), 5-11.
<https://doi.org/10.3102/0013189X018002005>
- Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111-130. <https://doi.org/10.21500/20112084.857>
- Milyavskaya, M., Gingras, I., Mageau, G. A., Koestner, R., Gagnon, H., Fang, J., & Boiché, J. (2009). Balance across contexts: Importance of balanced need satisfaction across various life domains. *Personality and Social Psychology Bulletin*, 35(8), 1031-1045.
<https://doi.org/10.1177/0146167209337036>

- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2), 246-268. <https://doi.org/10.1037/0033-295X.102.2.246>
- Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology*, 49, 229-258. <https://doi.org/10.1146/ANNUREV.PSYCH.49.1.229>
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton. <https://doi.org/10.1515/9783110813203>
- Molenaar, I., de Mooij, S., Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, 139, 107540. <https://doi.org/10.1016/j.chb.2022.107540>
- Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, 3(1), 2-20. <https://doi.org/10.3390/JINTELLIGENCE3010002>
- Morin, A. J. S. (2023). Exploratory structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 503-524). Guilford.
- Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1), 116-139. <https://doi.org/10.1080/10705511.2014.961800>
- Morin, A. J. S., Blais, A. R., & Chénard-Poirier, L. A. (2022). Doubly latent multilevel procedures for organizational assessment and prediction. *Journal of Business and Psychology*, 37(1), 47-72. <https://doi.org/10.1007/S10869-021-09736-5>
- Morin, A. J., Boudrias, J. S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., Madore, I., & Litalien, D. (2017). Complementary variable-and person-centered approaches to the dimensionality of psychometric constructs: Application to psychological wellbeing at work. *Journal of Business and Psychology*, 32, 395-419. <https://doi.org/10.1007/s10869-016-9448-7>
- Morin, A. J., & Marsh, H. W. (2015). Disentangling shape from level effects in person-centered analyses: An illustration based on university teachers' multidimensional profiles of effectiveness. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 39-59. <https://doi.org/10.1080/10705511.2014.919825>

- Morin, A. J., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (pp. 1044-1073). Wiley. <https://doi.org/10.1002/9781119568124.ch51>
- Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2022). Exploratory Structural Equation Modeling. In P. Fayers & R. Hays (Eds.), *The SAGE Encyclopedia of Research Design*. Sage Publications.
- Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41(5), 407-422. <https://doi.org/10.1016/J.INTELL.2013.06.004>
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Sage.
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Sage Publications. <https://doi.org/10.4135/9781412985413>
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74-98. <https://doi.org/10.1016/J.EDUREV.2017.08.004>
- Panadero, E., Jonsson, A., & Strijbos, J. W. (2016). Scaffolding self-regulated learning through self-assessment and peer assessment: Guidelines for classroom implementation. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 311-326). Springer. https://doi.org/10.1007/978-3-319-39211-0_18
- Paris, S. G., & Paris, A. H. (2003). Classroom applications of research on self-regulated learning. *Educational Psychologist*, 38(2), 89-101. https://doi.org/10.1207/S15326985EP3802_4
- Pedler, M. L., Willis, R. L., & Nieuwoudt, J. E. (2021). A sense of belonging at university: student retention, motivation and enjoyment. *Journal of Further and Higher Education*, 46(3), 397-408. <https://doi.org/10.1080/0309877X.2021.1955844>
- Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 259-282). Springer. https://doi.org/10.1007/978-1-4614-2018-7_12
- Perry, N. E. (1998). Young children's self-regulated learning and contexts that support it. *Journal of Educational Psychology*, 90(4), 715-729. <https://doi.org/10.1037/0022-0663.90.4.715>

- Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review*, 18(3), 211-228. <https://doi.org/10.1007/S10648-006-9014-3>
- Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review*, 16, 385-407. <https://doi.org/10.1007/S10648-004-0006-X>
- Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33-40. <https://doi.org/10.1037/0022-0663.82.1.33>
- Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, 53(3), 801-813. <https://doi.org/10.1177/0013164493053003024>
- Prokofieva, M., Zarate, D., Parker, A., Palikara, O., & Stavropoulos, V. (2023). Exploratory structural equation modeling: a streamlined step by step approach using the R Project software. *BMC Psychiatry*, 23(1), 546. <https://doi.org/10.1186/s12888-023-05028-9>
- Putarek, V., & Pavlin-Bernardić, N. (2020). The role of self-efficacy for self-regulated learning, achievement goals, and engagement in academic cheating. *European Journal of Psychology of Education*, 35(3), 647-671. <https://doi.org/10.1007/s10212-019-00443-7>
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69(2), 167-190. <https://doi.org/10.1007/BF02295939>
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. <https://doi.org/10.4324/9780203841624>
- Raykov, T., & Marcoulides, G. A. (2018). On studying common factor dominance and approximate unidimensionality in multicomponent measuring instruments with discrete items. *Educational and Psychological Measurement*, 78(3), 504-516. <https://doi.org/10.1177/0013164416678650>
- Reckase, M. D. (2009). Multidimensional Item Response Theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 607-642). Elsevier. <https://doi.org/10.1016/S0169-7161%2806%2926018-8>
- Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 183-203). University of Rochester Press.

- Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educational Psychologist*, 44(3), 159-175. <https://doi.org/10.1080/00461520903028990>
- Reeve, J. (2011). Teaching in ways that support students' autonomy. In D. Mashek & E. Y. Hammer (Eds.), *Enhancing teaching and learning: A leadership guide for school psychology* (pp. 90-103). Wiley. <https://doi.org/10.1002/9781444395341.CH5>
- Reeve, J., & Cheon, S. H. (2021). Autonomy-supportive teaching: Its malleability, benefits, and potential to improve educational practice. *Educational Psychologist*, 56(1), 54-77. <https://doi.org/10.1080/00461520.2020.1862657>
- Reeve, J., & Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology*, 98(1), 209-218. <https://doi.org/10.1037/0022-0663.98.1.209>
- Reeve, J., Jang, H., Hardre, P., & Omura, M. (2002). Providing a rationale in an autonomy-supportive way as a strategy to motivate others during an uninteresting activity. *Motivation and Emotion*, 26(3), 183-207. <https://doi.org/10.1023/A%3A1021711629417>
- Reeve, J., & Tseng, C.-M. (2011). Cortisol reactivity to a teacher's motivating style: The biology of being controlled versus supporting autonomy. *Motivation and Emotion*, 35(1), 63-74. <https://doi.org/10.1007/S11031-011-9204-2>
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696. <https://doi.org/10.1080/00273171.2012.715555>
- Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95(2), 129-140. <https://doi.org/10.1080/00223891.2012.725437>
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(1), 19-31. <https://doi.org/10.1007/s11136-007-9183-7>
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, 114(3), 552-566. <https://doi.org/10.1037/0033-2909.114.3.552>
- Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research*, 54(3), 429-443. <https://doi.org/10.1080/00273171.2018.1535420>
- Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, 23(1), 51-67. https://doi.org/10.1207/s15327906mbr2301_3

- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137-150.
<https://doi.org/10.1037/met0000045>
- Rosenzweig, E. Q., Wigfield, A., & Eccles, J. S. (2019). Expectancy-value theory and its relevance for student motivation and learning. In K. A. Renninger & S. E. Hidi (Eds.), *The Cambridge Handbook of Motivation and Learning* (pp. 617-644). Cambridge University Press. <https://doi.org/10.1017/9781316823279.026>
- Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, 66(1), 63-84.
<https://doi.org/10.1177/0013164404273942>
- Rusk, R. D. (2022). An Adaptive Motivation Approach to Understanding the 'How' and 'Why' of Wellbeing. *International Journal of Environmental Research and Public Health*, 19(19), 12784. <https://doi.org/10.3390/ijerph191912784>
- Rust, J., & Golombok, S. (2014). *Modern psychometrics: The science of psychological assessment*. Routledge. <https://doi.org/10.4324/9781315787527>
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1), 54-67.
<https://doi.org/10.1006/CEPS.1999.1020>
- Ryff, C. D. (2013). Eudaimonic well-being and health: Mapping consequences of self-realization. In A. S. Waterman (Ed.), *The best within us: Positive psychology perspectives on eudaimonia* (pp. 77-98). American Psychological Association.
<https://doi.org/10.1037/14092-005>
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). Springer.
https://doi.org/10.1007/978-1-4757-2691-6_5
- Samejima, F. (2016). Graded response models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 95-107). Chapman and Hall/CRC.
- Schmitz, B., & Schmidt, J. K. M. (2011). Assessing self-regulated learning using diary measures with university students: Technische universität darmstadt, darmstadt, germany. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 265-280). Routledge. <https://doi.org/10.4324/9780203839010.CH16>
- Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 1-15). Routledge.
<https://doi.org/10.4324/9781315697048-1>

- Schunk, D. H., Meece, J. L., & Pintrich, P. R. (2013). *Motivation in education: Theory, research, and applications*. Pearson.
- Shoda, Y., LeeTiernan, S., & Mischel, W. (2002). Personality as a Dynamical System: Emergence of Stability and Distinctiveness from Intra and Interpersonal Interactions. *Personality and Social Psychology Review*, 6(4), 316-325.
https://doi.org/10.1207/S15327957PSPR0604_06
- Shoda, Y., Mischel, W., & Wright, J. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, 67(4), 674-687.
<https://doi.org/10.1037/0022-3514.67.4.674>
- Silvestrin, M. (2024, August 29). Exploratory structural equation modeling in R.
<https://msilvestrin.me/post/esem/>
- Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, 25(7), E34-E46. <https://doi.org/10.1002/da.20432>
- Steel, P. (2007). The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological bulletin*, 133 1, 65-94.
<https://doi.org/10.1037/0033-2909.133.1.65>
- Steel, P., & Klingsieck, K. B. (2016). Academic Procrastination: Psychological Antecedents Revisited. *Australian Psychologist*, 51(1), 36-46. <https://doi.org/10.1111/AP.12173>
- Svartdal, F., Sæle, R. G., Dahl, T. I., Nemtcan, E., & Gamst-Klaussen, T. (2022). Study habits and procrastination: The role of academic self-efficacy. *Scandinavian Journal of Educational Research*, 66(7), 1141-1160. <https://doi.org/10.1080/00313831.2021.1959393>
- Tarka, P. (2017). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, 52, 313-354. <https://doi.org/10.1007/s11135-017-0469-8>
- Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses. *Applied Psychological Measurement*, 19(1), 39-49. <https://doi.org/10.1177/014662169501900105>
- Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, 60, 41-63. <https://doi.org/10.1016/j.jsp.2016.11.001>
- Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, 54(2), 119-135.
<https://doi.org/10.1016/j.erap.2003.12.004>

- Van Zyl, L. E., & Ten Klooster, P. M. (2022). Exploratory structural equation modeling: Practical guidelines and tutorial with a convenient online tool for Mplus. *Frontiers in Psychiatry*, 12, 795672. <https://doi.org/10.3389/fpsyt.2021.795672>
- Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, 3(1), 4-70. <https://doi.org/10.1177/109442810031002>
- Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The Sage Encyclopedia of Social Sciences Research Methods* (pp. 549-553). Sage Publications.
- Wei, X., Huang, J., Zhang, L., Pan, D., & Pan, J. (2022). Evaluation and comparison of SEM, ESEM, and BSEM in estimating structural models with potentially unknown cross-loadings. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(3), 327-338. <https://doi.org/10.1080/10705511.2021.2006664>
- Weinstein, C. E., Palmer, D. R., & Schultz, A. C. (2002). *LASSI: User's manual for those administering the Learning and Study Strategies Inventory*. H&H Publishing. <https://doi.org/10.3102/00028312023004614>
- Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, 55(2), 788-806. <https://doi.org/10.3758/s13428-022-01849-w>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, 25(1), 68-81. <https://doi.org/10.1006/CEPS.1999.1015>
- Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, 8(4), 327-353. <https://doi.org/10.1016/S1041-6080%2896%2990022-9>
- Winne, P. H. (1997). Experimenting to bootstrap self-regulated learning. *Journal of Educational Psychology*, 89(3), 397-410. <https://doi.org/10.1037/0022-0663.89.3.397>
- Winne, P. H. (2004). Comments on motivation in real-life, dynamic, and interactive learning environments: Theoretical and methodological challenges when researching motivation in context. *European Psychologist*, 9(4), 257-263. <https://doi.org/10.1027/1016-9040.9.4.257>
- Winne, P. H. (2005). A perspective on state-of-the-art research on self-regulated learning. *Instructional Science*, 33, 559-565. <https://doi.org/10.1007/S11251-005-1280-9>
- Winne, P. H. (2010a). Bootstrapping learner's self-regulated learning. *Psychological Test and Assessment Modeling*, 52(4), 472-490.

- Winne, P. H. (2010b). Improving measurement of self-regulated learning. *Educational Psychologist*, 45(4), 267-276.
- Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 29-46). Routledge.
- Winne, P. H. (2013). Self-regulated learning viewed from models of information processing. In D. H. Schunk & J. A. Greene (Eds.), *Self-regulated learning and academic achievement* (pp. 145-178). Springer. <https://doi.org/10.4324/9781410601032-9>
- Winne, P. H. (2017a). Cognition and metacognition within self-regulated learning. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 36-48). Routledge. <https://doi.org/10.4324/9781315697048.CH3>
- Winne, P. H. (2017b). Learning analytics for self-regulated learning. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 241-249). Society for Learning Analytics Research. <https://doi.org/10.18608/HLA17.021>
- Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, 112, 106457. <https://doi.org/10.1016/j.chb.2020.106457>
- Winne, P. H. (2021). Cognition, metacognition, and self-regulated learning. In D. S. Dunn (Ed.), *Oxford research encyclopedia of education*. Oxford University Press. <https://doi.org/10.1093/acrefore%2F9780190264093.013.1528>
- Winne, P. H. (2022). Modeling self-regulated learning as learners doing learning science: How trace data and learning analytics help develop skills for self-regulated learning. *Metacognition and Learning*, 17(3), 773-791. <https://doi.org/10.1007/s11409-022-09305-y>
- Winne, P. H., & Azevedo, R. (2022). Metacognition and self-regulated learning. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (3rd ed., pp. 93-113). Cambridge University Press. <https://doi.org/10.1017/9781108888295.007>
- Winne, P. H., & Baker, R. S. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining*, 5(1), 1-8. <https://doi.org/10.5281/ZENODO.3554619>
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9781410602350-19>

- Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research and applications* (pp. 297-314). Lawrence Erlbaum Associates.
<https://doi.org/10.4324/9780203831076-17>
- Winne, P. H., & Hadwin, A. F. (2013). nStudy: Tracing and supporting self-regulated learning in the Internet. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 293-308). Springer. https://doi.org/10.1007/978-1-4419-5546-3_20
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531-566). Academic Press.
<https://doi.org/10.1016/B978-012109890-2/50045-7>
- Winne, P. H., Jamieson-Noel, D., & Muis, K. (2002). Methodological issues and advances in researching tactics, strategies, and self-regulated learning. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement: New directions in measures and methods* (Vol. 12, pp. 121-155). JAI Press.
- Wolters, C. A., & Brady, A. C. (2020). College Students' Time Management: a Self-Regulated Learning Perspective. *Educational Psychology Review*, 33, 1319-1351.
<https://doi.org/10.1007/s10648-020-09519-z>
- Won, S., & Yu, S. L. (2018). Relations of perceived parental autonomy support and control with adolescents' academic time management and procrastination. *Learning and Individual Differences*, 61, 205-215. <https://doi.org/10.1016/j.lindif.2017.12.001>
- Won, S., Hensley, L. C., & Wolters, C. A. (2021). Brief research report: Sense of belonging and academic help-seeking as self-regulated learning. *The Journal of Experimental Education*, 89(1), 112-124. <https://doi.org/10.1080/00220973.2019.1703095>
- Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of Educational Psychology*, 116(1), 48-75.
<https://doi.org/10.1037/edu0000833>
- Woods, C. M., & Grimm, K. J. (2011). Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models. *Applied Psychological Measurement*, 35(5), 339-361. <https://doi.org/10.1177/0146621611405984>
- Wrzus, C. (2020). Experience sampling methods. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1483-1486). Springer.
https://doi.org/10.1007/978-3-319-28099-8_1300-1

- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100-1122. <https://doi.org/10.1177/1745691617693393>
- Yeager, D. S., & Dweck, C. S. (2020). What can be learned from growth mindset controversies? *American Psychologist*, 75(9), 1269-1284. <https://doi.org/10.1037/amp0000794>
- Zhang, B., Luo, J., Sun, T., Cao, M., & Drasgow, F. (2023). Small but nontrivial: A comparison of six strategies to handle cross-loadings in bifactor predictive models. *Multivariate Behavioral Research*, 58(1), 115-132. <https://doi.org/10.1080/00273171.2021.1957664>
- Zimmerman, B. J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology*, 25(1), 82-91. <https://doi.org/10.1006/CEPS.1999.1016>
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166-183. <https://doi.org/10.3102/0002831207312909>
- Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, 48(3), 135-147. <https://doi.org/10.1080/00461520.2013.794676>
- Zimmerman, B. J. (2015). Self-regulated learning: Theories, measures, and outcomes. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 541-546). Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.26060-1>
- Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, 80(3), 284-290. <https://doi.org/10.1037/0022-0663.80.3.284>
- Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, 53(1), 141-160.
- Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, 23(4), 614-628. <https://doi.org/10.3102/00028312023004614>
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123-133. <https://doi.org/10.1007/S11336-003-0974-7>

Appendix A: Study 1

Hadwin, A. F., Rostampour, R., Winne, P. H. (2024). Advancing Self-reports of Self-regulated Learning: Validating New Measures to Assess Students' Beliefs, Practices, and Challenges. Manuscript submitted for publication.

This research was supported by a Social Sciences and Humanities Research Council (SSHRC) of Canada insight research grant (PAR-IT; PI Allyson Hadwin, 435-2018-0440).

Abstract

Multi-modal research about self-regulated learning (SRL) has accelerated over the past few years. However, the development of self-report methods has not kept pace with other methodological advancements. The purpose of this study was to test the psychometric adequacy and predictive validity of four independent scales for measuring students' perceptions of SRL during studying. Participants were 473 post-secondary students across a range of academic disciplines. Confirmatory factor analyses confirmed the measurement adequacy of (a) a 4-factor SRL self-efficacy scale, (b) a 4 SRL importance scale, (c) an 8-factor self-regulated learning practices scale, and (d) a 6-factor academic challenges scale. The predictive validity of factors within each subscale revealed that (a) prioritizing and feeling confident about planning and foundational academic behaviours predicted academic performance, and (b) SRL practices were associated with either academic performance or academic challenges. Task understanding practices are underrepresented in most measures of SRL, but they proved to be important in this study predicting academic performance beyond other SRL practices. This instrument addresses a practical need in the field for measures that can be deployed in a range of classroom contexts by (a) teachers and researchers to test the effectiveness of SRL interventions and their impact on academic achievement (b) learners to identify targets for SRL improvement or development in subsequent study episodes (Sitzmann & Ely, 2011), and (c) institutions to conduct large scale assessment of SRL-related risk and protective factors for academic achievement.

Keywords: Self-regulated learning, SRL measurement, academic challenges, self-regulated learning strategies, self-efficacy for SRL, student success measures

Introduction

The past decade has witnessed a proliferation of research about self-regulated learning (SRL) coupled with important advancements in its measurement (c.f., Panadero et al., 2016). Most research has focused on refining the use of clickstream and multimodal data (Järvelä & Bannert, 2021; Winne, 2020b) for detecting traces of SRL practices as they emerge in situ (Biswas et al., 2018) and modelling traces of cyclical and sequenced events to examine the contingent nature of SRL (e.g., Bernacki, 2018; Järvelä et al., 2019). Unfortunately, the development of SRL self-report methods has not kept pace with these other methodological advancements.

We suggest there is value in developing more sophisticated self-report measures that provide insight into the metacognitive knowledge, beliefs and memories students use to make strategic decisions while building SRL competencies (Berger & Karabenick, 2016). Accessing learners' subjective experiences is essential for interpreting multi-modal traces of actions and calibrating those traces with metacognitive knowledge used to guide strategy deployment (Craig et al., 2020). We note three main limitations of classic self-report measures (e.g., LASSI, Weinstein et al. 1987; MSLQ, Pintrich et al. 1993) for validating and testing the calibration of observed-multi-modal indicators. First, they fail to collect data about why students choose to deploy specific tactics, actions, or practices during studying (Hadwin et al., 2001; McCardle & Hadwin, 2015, Winne, 2020). Second, they neglect to examine metacognitive beliefs and practices used for strategic decision making and adaptation during studying (McCardle et al., 2015; Tock & Moxley, 2017). Finally, they fail to provide a pragmatic framework for providing timely diagnostic feedback to learners and instructors. This paper contributes to a new wave of

research focused on pragmatic SRL measurements that concurrently serve both intervention and assessment purposes (Panadero et al., 2016).

Building on Winne and Hadwin's (1998) model of self-regulated learning, *the self-regulated learning profile and self-diagnostic instrument (SRL-PSD)* situates self-assessments of the self-regulatory actions (we refer to these as SRL practices) in the context of (a) beliefs students hold about their capabilities for successfully carrying out those practices (SRL self-efficacy), (b) perceptions of the importance of those practices (SRL importance), and (c) the academic challenges reported during academic work. The measure described in this paper is specifically designed to serve dual purposes as intervention and assessment by (a) prompting students to monitor and report on critical aspects of their self-regulated learning beliefs, practices and challenges, (b) producing an individualized self-diagnostic dashboard report for students, and, relevant to this study, (c) providing data for researchers to contextualize student perceptions about their SRL strategies (i.e., practices) in the context of their SRL beliefs and experiences with academic challenges. Rather than pitting self-reports against multi-modal or behavioral indicators of self-regulated learning, we advocate for advancing self-report measures as essential sources of information about the beliefs, actions and experiences learners perceive to be important during their studying that can inform interpretation of other multimodal behavioural and physiological indicators (Berger & Karabenick, 2016)

This instrument addresses a practical need in the field for measures that can be deployed in a range of classroom contexts by (a) teachers and researchers to test the effectiveness of SRL interventions and their impact on academic achievement (b) learners to identify targets for SRL improvement or development in subsequent study episodes (Sitzmann & Ely, 2011), and (c) institutions to conduct large scale assessment of self-regulatory engagement and beliefs learners

that predict academic outcomes (Schellings & vanHout-Wolters, 2011). The SRL-PSD instrument more fully represents self-report accounts of SRL by incorporating four independent scales (a) SRL self-efficacy, (b) SRL importance, (c) challenges experienced during studying), and (d) metacognitive SRL practices rarely accounted for in past measures of SRL (task understanding, metacognitive monitoring and evaluating, and metacognitive control or adaptation; McCardle et al., 2015).

The purpose of this study was to examine the psychometric properties and predictive validity of the SRL-PSD with a pilot and “test” sample of post-secondary students.

What is Self-Regulated Learning?

Self-regulated learning (SRL) is widely recognized as a critical competency for learners at all educational levels across the lifespan (Sitzmann & Ely, 2011, Zimmerman & Schunk 2011). A hallmark of SRL is learning to manage cognition, behavior, motivation, emotions, and manipulable factors in a learner’s environment (e.g., scheduling review, selecting external resources) by choosing, deploying, and adapting strategies in academic situations (Corno, 1986; Pintrich, 2000; Winne, 2018). Both the transition to post-secondary education and recently accelerated transitions to online instruction are examples of new academic situations with potential to present new challenges to studying.

From a self-regulatory perspective, successful learning is cyclical, recursive, and intentional. During self-regulatory processes learners actively take control of learning by (a) engaging effective planning and goal setting based on accurate and complete task interpretations, (b) deploying and managing effective task engagement practices, and (c) leveraging knowledge garnered from past experiences to fine-tune and improve how learning is conducted over time (Boekaerts & Corno 2005; Winne & Hadwin 1998; Zimmerman 1986, 1989). Metacognitive

monitoring and evaluation fuel self-regulated learning by generating knowledge about types of self-regulatory practices, and the efficiency and effectiveness of those practices (Winne, 2018). Self-regulated learning practices are associated with academic performance, persistence, effort, and positive or proactive beliefs about learning (de Bruijin-Smolders et al., 2016; Schunk & DiBendetto, 2016; Sitzman & Ely, 2011).

Winne and Hadwin (1998) modeled self-regulated learning as a recursive process unfolding over four key phases during which learners continuously develop, revise, and adapt (phase 4): (a) perceptions of the task and its context (phase 1), (b) goals and standards shaping plans and used to monitor progress (phase 2), approaches and (c) strategies for task engagement (phase 3). Within learning episodes, self-regulating learners judge products of their engagement (e.g., accuracy of task perceptions, adequacy of goals, effectiveness of strategies) against standards they hold for those products. This monitoring is what drives small-scale adaptation and improvement within a learning episode as well as large-scale revisions across learning episodes.

SRL's dynamic and recursive nature makes it difficult to measure (Winne, 2010). Products or outcomes at any phase of SRL become conditions for future phases. For example, goals students strive toward guide strategy choices they make during learning. Therefore, goal properties and standards are the outcome of phase 2 (goal setting). Overall and evolutionary outcomes arising as work on learning tasks progress condition strategy choices during phase 3 (task enactment). Understanding the context and timing of these contingencies can be critical for empirically modelling their interplay and impact on academic outcomes. Similarly, challenges or difficulties that arise when self-regulated learning practices fail become outcomes of SRL, and when detected in situ can also serve as conditions stimulating self-regulated learning practices for future work. For example, struggling to set a goal for a specific learning episode might serve

as a cue to revisit and refine earlier perceptions of task parameters generated during phase 1, task understanding. How students perceive and make sense of learning situations, actions taken, and experiences is essential to revealing factors that drive future SRL practices and changes in those practices.

Why develop a new self-report measure of self-regulated learning?

Measuring self-regulatory competencies and practices has garnered much discussion over the past decade (Hadwin, 2021; Panadero et al., 2017; Winne, 2019; Winne & Perry, 2000; Wolters & Won, 2018). Panadero et al. (2016) described the evolution of SRL measurement as unfolding over three waves. Early approaches tended to examine SRL as a disposition by asking students to respond to self-report instruments focused on their general or aggregated approaches to academic work (Wave 1). This wave was characterized using classic and still widely used questionnaires (e.g., MSLQ, Pintrich et al., 1993; LASSI, Weinstein et al., 1987) and interview protocols (e.g., SRLIS: Zimmerman & Martinez-Pons, 1986).

In the 1990's, attempts to examine the complex and multi-faceted nature of SRL as a process unfolding over time within learning episodes spawned a second wave of SRL measurement (Wave 2). These process-based measures relied heavily on clickstream (Winne, 2020b) and multimodal data (cf., Järvelä & Bannert, 2021) to represent SRL as events within and across learning episodes. This movement resulted in rapid changes to research methods for detecting SRL practices inductively as they emerge in situ (Biswas et al., 2018) and modelling cyclical and sequenced event data to describe the contingent nature of SRL (e.g., Bernacki, 2018; Järvelä et al., 2019). This second wave was also coupled with a movement away from the use of self-report measures in favour of “objective” observations of SRL activity.

However, as Panadero et al., (2016) point out, a third wave of research focuses on measurement approaches that have practical value as both assessment and intervention (Wave 3). These approaches strive to support learners to monitor and develop self-regulatory competencies, while simultaneously providing snapshots of SRL for researchers, instructors, and policy makers about changes in SRL over course and programs.

Students' self-reports about SRL may not align with observed traces of their SRL and metacognitive activity (e.g., Winne & Jamieson-Noel, 2002). However, they do provide insight into beliefs and memories upon which students design SRL and forge strategic decisions. Self-reports also have a key role to play in understanding students' metacognitive knowledge, awareness and regulation patterns underlying a range of observable actions or events (Craig et al., 2020) as they develop over time (McCardle et al., 2015). Students self-regulate based on perceptions and beliefs they hold about themselves, their learning situations/contexts, and the strategies they deploy (Hadwin & Winne, 2012; Winne & Hadwin, 1998; Winne & Marzouk, 2019). Self-reports provide a unique window into those beliefs and perceptions, albeit removed and generalized from a singular event in which SRL is exercised. Self-reports also provide immediately accessible and interpretable information that can be used by learners and instructors who seek information to recognize risky study practices and guide alternative approaches to studying in the immediate future. Further, self-reports may stimulate metacognitive awareness about aspects of SRL, in relation to SRL intentions, beliefs and academic challenges that emerge during studying. We propose leveraging the self-diagnostic potential of self-reports can play a key role in promoting metacognitive awareness to help students recognize, seek out and intentionally deploy strategies available to them through various interventions and resources. Also, in a systematic review of measures of self-regulated learning, Roth et al., (2016) found that

SRL behaviours were over-emphasized in measures of SRL, while beliefs such as self-efficacy were under-represented.

Unlike other areas of theoretical and methodological work in the field of SRL, research incorporating self-report data to examine SRL interventions has decelerated. Self-report measures (e.g., LASSI, Weinstein et al., 1987; MSLQ, Pintrich et al., 1993) typically examine SRL as a stable disposition, prompting responders to aggregate across unspecified periods of time, and failing to situate SRL practices differentially according to varying contexts of academic challenges and tasks learners encountered during studying (Hadwin et al., 2001; McCardle & Hadwin, 2015; Winne, 2020b). Furthermore, the items and subscales comprising these measures provide limited information about the metacognitive knowledge and practices learners use to make strategic decisions and adaptations to their studying (McCardle & Hadwin, 2015; Tock & Moxley, 2017).

This paper contributes to a Panadero's third wave of research focused on pragmatic SRL self-report measures that have practical purpose for assessing SRL and potentially informing interventions by instructors and learners (Panadero et al., 2016). Furthermore, we designed the scale to situate self-reports about SRL practices in the context of (a) beliefs students hold about those SRL practices and (b) challenges they continue to experience during studying.

For assessment purposes, *the self-regulated learning profile and self-diagnostic instrument (SRL-PSD)* provides insight to students, instructors and researchers about self-regulatory conditions, practices, and challenges students perceive during academic work. Informed by McCardle and Hadwin (2015), the SRL-PSD incorporates metacognitive planning and control practices including task understanding, goal management, monitoring/evaluating, and adaptation. However, the SRL-PSD also moves beyond metacognitive practices to incorporate important

behavioural, motivational, and cognitive facets of SRL that shape engagement in academic tasks. Specifically, students report on their use of (a) motivation regulation practices including task value, self-efficacy, and motivational appraisals, as well as motivation regulation practices, (b) time management practices, and (c) social academic engagement practices (e.g. peer support and helpseeking) deployed during a recent study episode.

Given the cyclical nature of self-regulated learning, self-reporting about learning may direct a student's attention to learning practices and experiences that might otherwise be submerged or even ignored. We acknowledge that most self-report items on questionnaires promote metacognitive awareness about those practices and experiences. While this may be viewed as a limitation of self-report, we view the potential of self-report measures for stimulating metacognitive awareness as a strength (Berger & Karabenick, 2016; Schellings & van-Houte Wolters, 2011).

Finally, the SRL-PSD addresses a practical need in the field for measures that can be deployed in a range of classroom contexts by teachers or researchers without specialized knowledge required to customize items to tasks and domains. The scale is designed to be useful for testing the effectiveness of SRL interventions on processes students engage during independent studying. It is also designed to identify within-person targets for future SRL improvement or development (Sitzmann & Ely, 2011). We posit that researchers and learners rely on similar data to understand and experiment with self-regulated learning practices and intentions (Hadwin et al., 2019a; Butler, 2002; Winne et al, 2011; Winne, 2022).

What Self-regulated learning practices are important for student success?

We define self-regulatory practices as the cognitive, motivational, behavioral, and metacognitive actions students engage to manage academic studying episodes across tasks and

disciplines. The contribution of self-regulatory practices to post-secondary academic success has been well documented and summarized (Broadbent & Poon, 2015; Jansen et al., 2019). Findings from these systematic reviews point to the complex influences SRL practices have on the difficulties students encounter during studying as well as success outcomes. Research points to the importance of at least five classes of SRL practices for academic performance outcomes including: task perceptions, goal setting and planning, motivation, metacognitive monitoring, and adaptation, and allocating and controlling time.

Task perceptions (forethought and planning) have been associated with academic beliefs such as self-efficacy (Miller, 2009), as well as academic performance and success outcomes (Callan & Cleary, 2019; Hadwin et al., 2008; Miller, 2009; Oshige et al. 2007). Beyond academic performance, task perceptions have been associated with other types of SRL practices such as goal and planning quality (Beckman et al., 2021; Greene et al., 2012), and by extension procrastination and disorganization (Cosnefroy et al., 2018). Hadwin and Winne (2012) proposed task perceptions construct meaning about three types of information including (a) explicit information about the task itself (e.g. instructions, page limits, etc.), (b) implicit information that can reasonably be inferred about the task (e.g., purpose, relation to other course components and activities), and (c) socio-contextual information related to the disciplinary meaning of such as task (e.g., engineers document plans and processes in ways that can be interpreted by others' years in the future).

Goal setting and planning practices have been found to correlate with post-secondary academic achievement (Jansen et al 2019). Students who set autonomous goals for attending university have higher academic performance and higher levels of both intrinsic and extrinsic motivation (Conti, 2000). Furthermore, students who list specific academic goals at the

beginning of a semester are more likely to have higher academic achievement (Acee et al., 2012). Finally, pursuing *mastery* goals is important for metacognitive monitoring (Jo et al., 2021), and instrumental help-seeking strategies (Howell & Buro, 2009; Roussel et al., 2011). However, self-reports about goal setting have tended to focus on learners' goal orientations (e.g., Durik et al., 2009; Zhou & Kam, 2017) rather than specific properties of goals (McCardle et al., 2017) or strategies and practices used to set and manage goals for optimizing learning episodes (see McCardle & Hadwin, 2015; McCardle et al., 2017).

Motivational practices. Motivational beliefs have been well-researched in the SRL and academic performance literatures (Fong et al., 2021). For example, intrinsic academic motivation, academic engagement, locus of control, and expectations for academic attainment, in combination with socio-emotional factors, play a key role in predicting high school retention and graduation. Expectancy for success and value beliefs predict academic choices and persistence, as well as academic performance (c.f., Eccles & Wigfield, 2002; Wigfield et al., 2016). Furthermore, interventions focused on improving utility-value beliefs are most effective for students with weaker pre-intervention academic performance (Hulleman et al., 2017; Rosenzweig et al., 2020). However, the ways students intentionally regulate maladaptive motivational beliefs during study episodes has been under-examined. We refer to these as practices used to regulate motivation. In this paper and the SRL-PSD, we intentionally distinguish between self-efficacy for self-regulation which is a motivational belief, and the strategies or practices students engage to control their self-efficacy.

Wolters (1998) originally identified motivation regulation strategies post-secondary students use during studying. Wolters and colleagues' self-assessments make a unique and important contribution to research on SRL because they shift the focus from motivational beliefs

and attributions that predict academic engagement to regulatory practices used to respond to maladaptive motivational beliefs during learning (e.g., Kim et al., 2020). Kim et al., (2020) found unique roles of three aspects of motivation (self-efficacy for self-regulated learning, task value, and regulation of motivation) in predicting achievement and behavioral outcomes associated with procrastination.

Self-efficacy enhancement strategies are approaches students use to control and enhance their self-efficacy. Proximal goal setting strategies (Zimmerman, 1990), metacognitive strategies such as self-monitoring, self-evaluation, and planning strategies (Bouchkioua, 2021; Zimmerman, 2000), seeking feedback, engaging cognitive strategies that promote mastery, and time management strategies have all been associated with enhanced self-efficacy judgements (Schunk & Ertmer, 2000)

The current study leverages this work by including practices used to regulate motivation during studying such as self-appraisals of motivational beliefs and task value. We also note these motivational practices share strong theoretical connections with task understanding, because they serve to focus learners on appraising and developing deeper understanding of assigned work in terms of its purpose and context (Hadwin & Winne, 2012).

Metacognitive monitoring and adaptation. Theoretically, regulation engages learners in a series of contingent events driving strategic adaptation (Winne, 2018; Winne & Hadwin, 1998). When learners detect misalignment between goals and progress, they are faced with several options: (a) persist with whatever they were doing and hope it will work better in the future, (b) try a new strategy or approach, (c) adjust or fine tune the current strategy, (d) update or change planning in the form of task understanding or goals, or (e) reduce effort or withdraw from the task altogether. Responding to previously detected problems invites a new round of monitoring

and evaluating which may then lead to continued refinement in approaches. This cyclical process is the essence of strategic and adaptive regulation. Despite the theoretical importance of adaptation and metacognitive control in SRL, student success research has virtually ignored this construct. While metacognitive practices including monitoring, self-evaluation and metacognitive awareness have been found to contribute to academic achievement (e.g., Callan & Cleary, 2019; Colthorpe et al., 2018), practices associated with adapting learning and strategic decision making as the result of monitoring and evaluation have been under-examined (Craig et al., 2020). Theoretically, adaptation practices deployed when strategies and approaches fail are essential to SRL (Butler & Winne, 1995; Hadwin & Winne, 2012).

Practices for *allocating and controlling time (time management)* have consistently been empirically linked to academic outcomes such as GPA (Thibodeaux et al., 2017; Wolters & Brady, 2020), particularly when those practices leveraged goal setting and prioritization of task and goals (Adams & Blair, 2019). Recent research indicates pre-university time management and autonomous motivation positively predict academic achievement in the first year of post-secondary studies (Marjon Fokkens-Buinsma et al., 2021). Finally, in a comprehensive review of the literature, Wolters and Brady (2020) provide strong theoretical framework for understanding the role of time management as a self-regulatory process contributing to a range of academic outcomes including well-being, motivation, and procrastination in college students.

Findings in the student success literature consistently report the importance of *social and emotional factors* for academic learning and performance as students transition to university (Tinto, 1982; Wong et al., 2024). While social factors are amongst a set of protective factors for successful university transitions (see Ball et al., 2024), the focus tends to be on contextual and environmental factors such as social support, and school climate. The student success literature

consistently documents the importance of social adjustment, social integration (Veldmen et al., 2019), social belongingness (Baumeister et al., 2017), and social well-being (Keyes, 2005; Knoesen & Naudé, 2017). From a self-regulated learning perspective, we are also interested in the things students do to exercise agency and control of these social factors during their studying. However, a paucity of research focuses on the things students do to leverage their social adjustment, integration, and perceptions of belongingness during academic work. A review of items on post-secondary student engagement surveys (e.g., NSSE, OSE, SCEQ) revealed at least three ways students exercise agency in their academic social engagement during coursework including: working with peers to provide academic assistance to one another, leveraging and enjoying opportunities to engage in discussions and interactions with peers and instructors, and making friends in their classes (Dixson, 2015; Handelsman et al, 2005; Kuh, 2009).

Why examine self-regulated learning beliefs and intentions as conditions for deploying self-regulatory practices?

Beliefs students hold about their self-regulatory competencies are important conditions shaping goals for learning, intent to self-regulate, and strategies deployed during regulation (Trautner & Schwinger, 2020; Wolters & Benzon, 2013). We identify two types of beliefs that contextualize deliberate choices students make during self-regulation.

Self-efficacy for self-regulated learning refers to belief's learners hold about whether they can successfully self-regulate learning (Usher & Pajares, 2008). These beliefs have been strongly associated with productively deploying strategies (Bandura et al., 1996; 2001), help-seeking (Won et al., 2021), time management and procrastination (Klassen et al., 2010; Won & Yu, 2018), behavioural engagement (Putarek & Pavlin-Bernardić, 2020) academic performance (Klassen, 2020), and intentions to persist at college (Her & Thompson, 2022). Self-efficacy for

self-regulation scale asks students to rate their beliefs about their capability to deploy specific self-regulation activities (e.g., Zimmerman et al., 1992). One limitation of these measures is their focus on global judgements of capability (e.g., How well you can...), rather than judgements of capability for a specific task or study period (e.g., during this task, or in the next two weeks). This omission fails to distinguish between self-efficacy beliefs about potential or future self, versus self-efficacy beliefs for a proximal task or goal (Bandura, 1997). A second limitation of scales used in these studies is the focus on self-efficacy for specific skills (e.g., taking notes, organizing schoolwork, remembering information) rather than efficacy for deploying specific regulatory processes (e.g., understanding what they need to do, managing goals, monitoring learning). Naming the specific skill conflates SRL with deploying study skills and underrepresents the strategic and metacognitive nature of SRL.

SRL Importance. Intentionality is a hallmark of SRL. Theoretically, intent to purposefully take control of learning in combination with self-efficacy should predict academic engagement and success in those endeavours (Winne, 2022). Intention to self-regulate learning has not been given much empirical attention to date. However, we posit that when SRL practices are not viewed as important, they are unlikely to be prioritized or meaningfully deployed during studying. In other words, students are unlikely to engage in SRL practices if they are not considered to be important for their academic success.

Why examine challenges experienced during the deployment of self-regulated learning practices?

Theoretically, challenges related to academic tasks, as well as perceptions that aspects of studying are not optimal, create opportunities to exercise self-regulatory control in the future (Hadwin et al., 2018; 2019). In other words, SRL practices can be regulatory responses to academic challenges, potentially buffering the deleterious impact of those challenges on

academic performance. The SRL-PSD invites students to report on challenges or struggles that characterized their recent studying attempts. These challenges serve as a backdrop for (a) interpreting the effectiveness of strategies or practices used during those study periods and (b) thinking forward to strategies or practices that might be better suited to those challenges in future. Engaging in this type of reflection about studying focuses on leveraging opportunities for finetuning and improving learning experiences.

Despite extensive research about the factors contributing to student success, scant research examines the specific challenges students report during studying. Findings to date indicate time management, cognitive strategies, concentration, and tiredness are the challenges most frequently reported by learners (Koivuniemi, et al., 2017). Koivuniemi and colleagues also found students who reported elevated levels of generalized SRL skills (as represented by the MSLQ) reported a higher frequency of cognitive challenges during individual learning than students ranking low on SRL skills. Similarly, Hadwin et al., (2019) found the most frequently reported weekly challenges over an academic semester included: motivation and emotion, goal setting and planning, and cognitive challenges. Motivation and emotional challenges were consistently high, regardless of level of goal attainment. Consistent with findings reported by Koivuniemi, et al., (2017), cognitive challenges tended to be more frequently reported among students who were academically successful, thereby attaining goals they set for studying that week.

Purpose

Drawing from the theoretical and empirical foundations described above, this study introduces the SRL Profile and Self-Diagnostic instrument (SRL-PSD). The SRL-PSD incorporates separate scales for (a) beliefs about SRL, (b) SRL practices, and (c) challenges students have encountered during their recent studying. The purpose of this study is to examine psychometric properties of the SRL-PSD with a sample of post-secondary university students. First, confirmatory factor analysis was used to test the factor structure of each of the four SRL-

PSD scales. Second, structural equation modelling was used to explore the predictive validity of these scales. Two research questions were examined:

(RQ1) Are the four scales psychometrically adequate?

(RQ2) Do the four scales have predictive validity for understanding academic outcomes?

Methods

Participants

Pilot Sample

A pilot sample of 220 university students from a wide range of programs and academic levels who participated in an introductory Learning to Learn course was used for initial piloting of the scales and conducting exploratory factor analyses (EFA) on the scales. Students responded to a battery of questionnaires for self-assessment and as part of their course requirements during week 10 of a 13-week academic semester. The students were from 8 faculties (e.g., education, social science, engineering, computer sciences, fine arts, humanities, business) and had a variety of academic performance levels, with an average GPA of 5.2 out of 9 which corresponds to the letter grade B (SD = 1.9, range = 1-9). 76% of students were first-year students, 16% second-year, 6% third year and 2% fourth year students. The sample included 89% full-time and 11% part-time students, with 48% being female. In terms of residency, 86% were Canadian citizens, 3% were permanent residents, and 11% were international students on study permits. The mean age of the students was 20.1 years (SD = 2.5, range = 18-49). This sample was used solely for the initial piloting of the scales; all analyses in the current study were conducted using the main sample.

Main Sample

Participants were undergraduate students ($N = 473$) enrolled in a first-year course at Western Canadian university (74% female) with a mean age of 20.4 (sd=3.3). They were predominately first-year (39%) Canadian citizens (89%) from the Faculty of Social Sciences

(55%), Faculty of Science (17%), or one of the other faculties (see Table 1). Participants were recruited through (a) the Psychology department's research participation pool ($n = 354$), (b) an elective educational psychology learning-to-learn course ($n = 88$), or (c) a digital learning strategy library available to students across the university ($n = 31$). The average GPA of participants was 6.6 out of 9, which corresponds to a letter grade between B+ and A- ($SD = 1.8$). Students from the psychology pool received one course credit to allocate to any of their psychology courses. Students enrolled in the elective course completed the questionnaires as part of course requirements and self-study activities. Students from the strategy library were enrolled in a draw for an amazon gift card. All participants gave informed consent to participate in research.

Table 1
Participants Demographics

<i>Faculty</i>		
Faculty of Social Sciences-Dep. Of Psychology	186	39%
Faculty of Social Sciences-Other	77	16%
Faculty of Science	82	17%
Faculty of Engineering	41	9%
Faculty of Humanities	32	7%
School of Business	21	4%
Faculty of Education	16	3%
Faculty of Fine Arts	12	3%
Faculty of Human & Social Development	5	1%
<i>Year of Study</i>		
First Year	184	39%
Other years	289	61%
<i>Citizenship Status</i>		
Canadian Citizens	422	89%
Permanant Residents	14	3%
International students on study permits	37	8%
Total	473	100%

Pandemic Context

Importantly, data for this study were collected in Fall 2020 during a unique and unprecedented pandemic. All students attended university through online off-campus education. Online instruction was new for most students and instructors. Students were geographically distributed across nations. Most students studied from home. In our regional context, COVID restrictions limited socializing beyond the family or a small cohort of 4-6 friends. Many extra-curricular activities were either cancelled or significantly restricted. Implications of this pandemic context for interpretations and generalizability of findings is taken up in the discussion.

The SRL-PSD Instrument

The SRL-PSD items and subscales were developed iteratively over a decade of data collection and pilot testing with post-secondary students. Further description about development is provided for each scale below. Specific items are presented in Appendix A. Students were prompted to form responses by aggregating experience over the past two recent weeks of studying. This was done to intentionally sample studying practices and challenges across a range of tasks and subject areas while reducing forgetting and unrepresentative leveling likely to arise when responses represent longer intervals (e.g., “*Rate the following items based on your behavior in this class*” MSLQ: Pintrich & DeGroot, 1990).

Self-Regulated Learning Practices. The SRL- Practices Scale (SRL-P) measures students’ self-reported engagement in actions aligned with SRL processes, herein referred to as SRL practices. The SRL-P comprises 31 items yielding 8 subscales related to (1) task understanding (e.g., “Asked myself if I know what is important to learn”), (2) goal management (e.g. “Set goals for my work”), (3) appraisal of task value (e.g. “Reflected on why this work is important”), (4)

motivation appraisal (e.g. “Assessed if think I can do it”), (5) time management (e.g., “Created a timeline or schedule), (6) metacognitive monitoring (e.g. “Asked myself if I am understanding the material”), (7) metacognitive adaptation (e.g. “Modified my plans for the task”), (8) academic social engagement (e.g. “Got to know people in the class”). Reporting on the two previous weeks, participants were required to rate their level of agreement about engaging each practice on a five-point Likert scale ranging from (1) strongly disagree to (5) strongly agree. High scores indicate more proactive SRL behaviours or practices.

Winne and Hadwin’s (1998) model of studying as self-regulated learning was used to prioritize item selection that would be sensitive to (a) metacognitive practices underrepresented in past measures related to task understanding, monitoring and evaluating and adapting, (b) strategic actions deployed during studying rather than task specific techniques, and (c) self-regulatory processes across phases of SRL. Specifically, items were selected by drawing from (a) an earlier regulated learning scale (RLQ; McCardle & Hadwin, 2015), (b) regulation of motivation items such as those used by Kim et al. (2020), (c) time management and procrastination items such as those used by (Won & Yu, 2018), and (d) social belongingness and engagement practices such as those examined by Won et al. (2021). The latter were added quite auspiciously to be sensitive to social and emotion-supporting practices that might be particularly important for students to think about during a period of pandemic driven online learning. Finally, items were compared to weekly open-ended descriptions of “strategies” used to confront academic challenges that constrained goal attainment or progress toward weekly self-set study goals.

Self-Regulated Learning Challenges. Academic challenge items were originally identified by examining open-ended responses to a weekly diary item asking students to describe

challenges affecting their progress toward self-set goals for studying. Over several years, challenges were reviewed and organized loosely into broad categories that aligned with five factors identified in the student success literature (e.g., Bowden et al., 2021; Wong et al., 2024): behavioural, motivational, cognitive, metacognitive, and social and emotional challenges. Saturation was satisfied when no new challenges were identified over two subsequent semesters. Those challenges became the items used in the SRL-PSD scale.

The SRL-C (Self-Regulated Learning Challenges scale) comprises 31 items (6 subscales) assessing the degree to which students encountered a range of challenges in their studying. The *Motivation challenge* subscale is comprised of 3-items related to motivational beliefs, interest, and persistence. The *Metacognitive challenge* subscale includes 8-items related to task understanding, goal setting and planning, and monitoring. The *Cognitive challenge* subscale has 6-items related to attending, encoding, and remembering. The *Behavioral challenge* subscale is comprised of 8-items related to managing time, tasks, and procrastination. The *Socio-emotional challenge* subscale has 6-items related to emotional, social, and relational aspects of academic success. Reporting on the two previous weeks, participants were required to rate their level of agreement about experiencing each challenge on a five-point Likert scale ranging from (1) strongly disagree to (5) strongly agree. High scores indicate a student is struggling to manage aspects of studying theoretically and empirically associated with student success and performance.

SRL Importance. The SRL Importance scale (SRL-I) includes 10 items describing broad categories of SRL practices (e.g., understanding what I need to do; managing my time, etc.). Selective items were chosen to be most representative of items in the SRL-practices subscale. Students rated the importance of each item for their studying in the last 2 weeks on a 5-point

Likert scale from (1) very low priority to (5) very high priority. High scores indicate students view those SRL practices as important priorities during studying.

SRL Efficacy the SRL self-efficacy scale (SRL-SE) includes the same 10 items from the SRL Importance scale. This time students were asked to rate their self-efficacy for successfully engaging those SRL practices over their next two weeks of studying on a 5-point Likert scale from (1) not confident to (5) completely confident. High scores indicate students have high self-efficacy or confidence they can successfully engage SRL practices during future studying.

Academic Performance

Academic performance was measured using end of semester registry GPA gathered from institutional records. Semester GPA represents performance across all courses in that academic semester and is reported on a 9-point scale where 0 = F(0-48%), 1 = D (50-59%), 2 = C(60-64%), 3 = C+(65-69%), 4 = B-(70-72%), 5 = B(73-76%), 6 = B+(77-79%), 7 = A-(80-84%), 8 = A(85-89%), and 9 = A+(90-100%).

Procedures

The SRL-PSD was administered in week 11 of a 13-week academic semester in the Fall of 2020. Items within each scale were randomly presented online in LimeSurvey (V3.2, 2020).

Analytic Strategies

Ranges for skewness range of ± 1.5 and kurtosis of ± 2 for items were regarded as indication of univariate normality in the data (Byrne, 2010; Hair et al., 2010; Kline, 2015). All statistical analyses were done in R (lavaan package (v0.6-12); Rosseel, 2012) using a robust maximum likelihood estimator (MLR). Compared to maximum likelihood estimator (ML), MLR is less sensitive to outliers and to deviations from multivariate normality as it corrects for non-normality-induced bias in the standard errors (Finney & DiStefano 2008). There were no missing

data; participants responded to the full survey in a single session and university recorded Registry GPA data were available for all participants.

Confirmatory factor analysis (CFA) was used to examine *a priori* factor structure of the scales. The goodness of global and approximate fit was regarded acceptable using CMIN/DF<3; CFI and TLI >.9, RMSEA and SRMR <.08 criteria (Byrne 2001; Hu & Bentler, 1999; Kline 2015; Schumacker & Lomax, 2004). The criteria for evaluating the fit indices are summarized in Table 2.

Table 2 <i>Evaluation Criteria for Interpreting the Fit Indices</i>	
Indices	Evaluation Criteria
Normed Chi-Square (χ^2/df)	≤ 5 indicates an acceptable fit. ≤ 3 indicates a good fit.
Comparative Fit Index (CFI)	≥ 0.90 indicates an acceptable fit ≥ 0.95 indicates a good fit
Tucker-Lewis Index (TLI)	≥ 0.90 indicates an acceptable fit ≥ 0.95 indicates a good fit
Root Mean Square Error of Approximation (RMSEA)	0.06-0.08 indicates an acceptable fit 90% CI< 0.08 indicates an acceptable fit p value for H_0 RMSEA ≥ 0.080 be $<.05$
	0.01-0.05 indicates a good fit 90% CI< 0.05 indicates a good fit p value for RMSEA ≤ 0.050 be $>.05$
Standardized Root Mean Square Residual (SRMR)	0.06-0.08 indicates an acceptable fit 0.01-0.05 indicates a good fit

Regarding size and significance of item-factor loadings, $B=0.5$ was regarded as the minimum acceptable amount of communality between items and their corresponding factors (Tabachnick & Fidell, 2007). To estimate the reliability of the factors, McDonald's Omegas (ω)

were calculated which are often preferred in the context of multidimensional scales with unequal factor loadings (Hayes & Coutts, 2020).

To test if a unidimensional structure fit the data for the SRL-Importance and SRL-Self Efficacy scales, we ran exploratory factor analysis (EFA) first using the pilot sample. We then validated the structure informed by the EFA using confirmatory factor analysis (CFA) with the current sample. Similarly, the *a priori* 5-factor structure for the academic challenges scale was not supported by the pilot study. Information on how to refine the structure was derived from conducting an EFA using the pilot sample, which then informed the CFA conducted with the current sample.

To test the predictive validity of the SRL-PSD, structural equation modelling (SEM) was used to regress end-of-semester GPA (dependent variable) on SRL-PSD subscales (independent variables). We tested predictive validity using two types of models (a) a series of bivariate models in which a single subscale predicted GPA independently, and (b) a series of models in which all subscales of a scale predicted GPA concurrently. The bivariate estimates can be interpreted as zero-order associations between the subscales and GPA. We opted for this strategy to understand the independent contribution of each subscale in predicting academic performance as well as to understand the additive and suppressive effects of subscales in simultaneous prediction of academic performance.

Results

Examining the psychometric adequacy of the 4 scales? (RQ1)

Descriptive statistics for all indicator variables and correlation matrix for the factors are presented in Appendix B (Table B2 & B3). Items significantly loaded onto their corresponding factors (Appendix B, Tables B5, B7, & B9). Regarding reliability, most subscales exceeded the

conventional cut-off $\omega = .700$ (Lance et al., 2006) except for *Foundational Academic Behaviours* scale in SRL-I and SRL-SE, *time management* and *motivation appraisal* in SRL-P and *motivation* challenges in SRL-C. These subscales had only three items for which ω ranged from .646 to .696. Reliability estimates are reported in Appendix B (Table B4).

Table 3

Model fit indices for each of the 4 scales

	SRL-I	SRL-SE	SRL-P	SRL-C
Baseline model X^2	1345.528	1742.190	3943.923	5926.728
Baseline model df	78	78	465	465
Factor model X^2	151.020	173.035	684.594	738.464
Factor model df	59	59	405	419
p	<.001	<.001	<.001	<.001
χ^2/df	2.56	2.93	1.69	1.76
CFI	0.927	0.931	0.924	0.942
TLI	0.904	0.909	0.913	0.935
RMSEA	0.057	0.064	0.042	0.040
RMSEA 90% CI	[.047,.068]	[.054,.074]	[.036,.047]	[.036,.045]
P-value H_0 :	0.115	0.011	0.994	1.000
RMSEA ≤ 0.050				
P-value H_0 :	0.000	0.004	0.000	0.000
RMSEA ≥ 0.080				
SRMR	0.059	0.062	0.050	0.048
Estimator	MLR	MLR	MLR	MLR

Note. Abbreviations: SRL-I: SRL-Importance; SRL-SE: SRL-Self-Efficacy; SRL-P: SRL

Practices; SRL-C: Academic Challenges

A 4-factor solution exhibited acceptable fit for SRL-I (importance) and SRL-SE (self-efficacy) scales (Table 3). All items significantly loaded onto one of the following factors:

Planning, Metacognition and Beliefs, Foundational Academic Behaviours and *Academic Social Engagement*. Item-factor loadings and covariance among factors for these scales are provided in Appendix B (Tables B5 and B6).

Findings also indicated an acceptable fit for an 8-factor solution for the Self-Regulated Learning Practices (SRL-P) scale including *Task Understanding*, *Goal Management*, *Time Management*, *Monitoring*, *Adaptation*, *Appraisals of Task Value*, *Motivation Appraisal*, and *Academic Social Engagement* (Table 3). All items adequately loaded on their corresponding factors. Two items from the goal management factor (“evaluated my progress toward goals” and “considered whether my goals are appropriate for the task at hand”) were identified as redundant and removed. The residual variance for a pair of items were allowed to covary. These items were “assessed if I think I can do it” (MA1) and “assessed my feelings for the task” (MA2) from *Motivation Appraisal* subscale ($r=.30$, $p<.001$). We assumed the residual covariance between these two items is due to a shared method effect (e.g., using the exact word “assessed” in both items). The factorial structure and parameter estimates for the SRL-P scale are shown in Appendix B (Table B7 and B8).

The *a priori* 5-factor solution for the *Academic Challenges* scale (SRL-C) included behavioral, motivational, cognitive, metacognitive, and social and emotional factors. However, a pilot study revealed that certain items did not align well with their *a priori* factors. An EFA on the pilot sample ($N=220$) proposed a revised 6-factor structure for the scale, which introduced a distinct factor for items related to *initiating and sustaining engagement* (e.g., “wanting to do schoolwork”, “putting things off”, “managing distractions”). Additionally, the remaining items that focused on behavioral challenges were grouped under a new factor named *Goal and Time Management challenges*. This 6-factor solution was tested using the main sample and demonstrated an acceptable fit, as detailed in Table 3. Items and parameter estimates are provided in Appendices B (Table B9 & B10).

The global fit of the measurement models for the 4 scales are summarized in Table 4. Translating numbers to language descriptors (i.e., acceptable, good) is based on the criteria provided in Table 2.

Table 4*Evaluation of Fit Indices for the scales*

Fit indices	SRL-I	SRL-SE	SRL-P	SRL-C
Normed Chi-Square (χ^2/df)	Good Fit	Good Fit	Good Fit	Good Fit
Comparative Fit Index (CFI)	Acceptable Fit	Acceptable Fit	Acceptable Fit	Acceptable Fit
Tucker-Lewis Index (TLI)	Acceptable Fit	Acceptable Fit	Acceptable Fit	Acceptable Fit
RMSEA	Good Fit	Acceptable Fit	Good Fit	Good Fit
SRMR	Good Fit	Acceptable Fit	Good Fit	Good Fit

Examining the Predictive Validity of the Four Scales for Explaining Academic Performance (RQ2)

Results of bivariate models in which SRL-I facets independently predicted GPA indicated that prioritizing *foundational academic behaviours* ($B = .61, p < .001$) and prioritizing planning practices ($B = .56, p < .001$) positively predicted end of semester GPA while the effects for the other two facets of this scale (MB-I and ASE-I) were statistically insignificant. When all four facets were included in model (Table 5, concurrent-predictor model), prioritizing *foundational academic behaviour* positively predicted GPA ($B=.84, p=.002$) while prioritizing practices related to *academic social engagement* negatively predicted GPA ($B=-.57, p=.006$). The regression coefficients, R^2 , and fit indices for this model are provided in Table 5.

Table 5*SRL-Importance scales predict end of semester GPA*

Predictors	Coef. (B)	
	Bivariate Associations with GPA	Concurrent-Predictor Model (outcome=GPA)*
Plan-I	0.559***	0.363
MB-I	0.018	-0.257
FAB-I	0.610***	0.841**
ASE-I	-0.247	-0.567**

Note. *Model syntax: Plan-I+MB-I+FAB-I+ASE-I → GPA

Fit indices: $X^2(68) = 166.6***$, CFI=.925, TLI=.9, RMSEA=.055 [.046, .065], SRMR= .059, $R^2 = .112$

Correlation among predictors can be found in the Appendix C, Table C1.

Abbreviations: SRL-Importance (-I): (Plan: Planning; MB: Metacognition and beliefs; FAB: Foundational academic behaviours; ASE: Academic Social Engagement).

We conducted a follow up analysis to understand the negative association between prioritizing academic social engagement (ASE-I) and GPA in the concurrent-predictor model.

After removing non-significant predictors from the model, there was no change the in the pattern, effect size, or global fit of the model. The correlation between the significant predictors in this model, namely, ASE-I and FAB-I was $r=.39, p<.001$. No statistically significant bivariate association between ASE-I and GPA was found ($B=0.247, p=0.085$). The concurrent-predictor model indicated that prioritizing higher academic social engagement (ASE-I) predicted lower academic performance when the effect of prioritizing foundational academic behavior (FAB-I) was partialled out.

Similarly, results from bivariate models in which SRL self-efficacy facets independently predicted GPA (Table 5) showed that confidence in being able to deploy *planning* practices ($B = .78, p = <.001$) and enact *foundational academic behaviours* ($B = .76, p = <.001$) predicted end of semester GPA while self-efficacy about *metacognition and beliefs*, and *academic social engagement* did not. When considering all subscales together in predicting GPA (Table 6; column 2), confidence in deploying *planning* practices was no longer a significant predictor of GPA but confidence in engaging in *foundational academic behaviours* continued to positively predict GPA ($B=.97, p<001$).

Table 6
SRL-Self-Efficacy scales predict end of semester GPA

Predictors	Coef. (B)	
	Bivariate Associations with GPA	Concurrent-Predictor Model (outcome:GPA)*
Plan-SE	0.776***	0.01
MB-SE	0.242	-0.104
FAB-SE	0.755***	0.966***
ASE-SE	0.078	-0.341

Note. * Model syntax: Plan-SE+MB-SE+FAB-SE+ASE-SE→GPA.

Fit indices: $\chi^2(68) = 189.7***$, CFI=.930, TLI=.906, RMSE=.062 [.052, .071], SRMR= .061, $R^2 = 0.115$

Correlation among predictors can be found in the Appendix C, Table C2.

Abbreviations: SRL-Self-Efficacy (SRL-SE): (Plan: Planning; MB: Metacognition and beliefs; FAB: Foundational academic behaviours; ASE: Academic Social Engagement).

For the SRL practices scale, results from bivariate models (Table 7) showed that *task understanding* practices ($B = .94, p < .001$), *goal and time management* practices ($B = .4, p < .001$), *Time Management* practices ($B = .57, p = .001$), and *Monitoring* practices were significant independent predictors of GPA. Higher reports of these practices were associated with higher GPA. However, motivational practices (*Appraisals of Task Value and Motivation appraisal*), *adapting*, and *academic social engagement* practices were not found to be statistically significant predictors of GPA. When GPA was simultaneously regressed on all SRL-P facets in the concurrent-predictor model, only *task understanding* practices predicted GPA ($B=1.47, p<.001$). Notably, *task understanding* practices were moderately to highly correlated with *goal management* ($r=.67, p<.001$), *time management* ($r=.54, p<.001$), and *monitoring practices* ($r=.53, p<.001$) (Appendix C, Table C3).

Table 7

SRL-Practices Scales Predict end of semester GPA

Predictors	Coef. (B)	
	Bivariate Associations with GPA	Concurrent-Predictor Model (outcome: GPA)*
Task Understanding	0.944***	1.471***
Goal Management	0.401**	-0.472
Appraisals of Task Value	0.1	0.151
Motivation Appraisal	0.241	-0.814
Time Management Practices	0.572**	0.598
Adaptation Practices	-0.285	-0.417
Monitoring Practices	0.337*	0.113
Academic Social Engagement	0.106	0.011

Note. *Model syntax: TU+GM+ATV+MA+TM+AD+MON+ASE→GPA

Fit: $X^2(427) = 712.4***$, CFI=.920, TLI=.907, RMSE=.038 [.033, .042], SRMR= .049, $R^2 = 0.128$

Correlation among predictors is provided in the Appendix C, Table C3.

Abbreviations: SRL-Practices (-P): (TU: Task Understanding; GM: Goal Management; ATV: Appraisals of Task Value; MA: Motivation Appraisal; TM: Time Management Practices; AD: Adaptation Practices; MON: Monitoring Practices; ASE: Academic Social Engagement).

Finally, the results of bivariate models where each SRL challenge facet (SRL-C) predicted GPA (Table 8) showed that *motivation* challenges ($B=-.53, p<.001$), *cognitive* challenges ($B=-.68, p<.001$), *metacognitive* challenges ($B=-.58, p<.001$), *initiating and sustaining engagement* challenges ($B=-.29, p=.01$) and *goal and time management* challenges ($B=-.46, p<.001$) all

significantly predicted GPA while *social and emotional* challenges did not. Such inverse associations indicated that higher levels of perceived challenges were associated with lower GPA. In the concurrent-predictor model (Table 8, column2), *motivation* challenges negatively predicted GPA ($B=-1.25, p=.02$) while *social and emotional* challenges positively predicted GPA ($B=1.1, p<.001$).

As *motivation* challenges were highly correlated with all the other factors in the SRL-C scales (average $r = .72$, see Appendix C, Table C4), multicollinearity was a concern. Therefore, the non-significant paths were removed from the concurrent-predictor model. In the parsimonious model, which only included two predictors (MOT-CH and S/EM-CH), the pattern of associations, global fit, and effect sizes ($B = -1.13, p<.001$ for MOT, and $B=.94, p<.001$ for SE-CH, $R^2=.15$) did not significantly change, indicating that the estimates for the model that included all concurrent predictors (Table 8, column2) were not biased due to multicollinearity.

The positive association between *social & emotional* challenges and GPA in the concurrent-predictor model can be better understood by comparing its bivariate association with GPA (Table 8). When the effect of other challenges (mainly *motivation* challenges) was controlled, higher levels of *social & emotional* challenges were associated with obtaining higher GPA scores. We return to this finding in the discussion section.

Table 8
Academic Challenges scales predicting end of semester GPA

Predictors	Coef. (B)	
	Bivariate Associations with GPA	Concurrent-Predictor Model (outcome: GPA)*
Motivation-CH	-0.530***	-1.445*
Cognitive-CH	-0.680***	0.187
Metacognitive-CH	-0.578***	-0.384
Initiating & Sustaining Eng.-CH	-0.287*	0.367
Goal & Time Management-CH	-0.459***	0.196
Social & Emotional-CH	-0.043	1.072***

Note. Model syntax: MOT-CH +COG-CH +MET-CH +ISE-CH+GTM-CH +S/EM-CH→GPA

Fit indices: $X^2(444) = 778.6^{***}$, CFI=.940, TLI=.933, RMSE=.04 [.036, .044], SRMR= .047, $R^2 = .184$

Correlations among predictors are provided in Appendix C, Table C4.

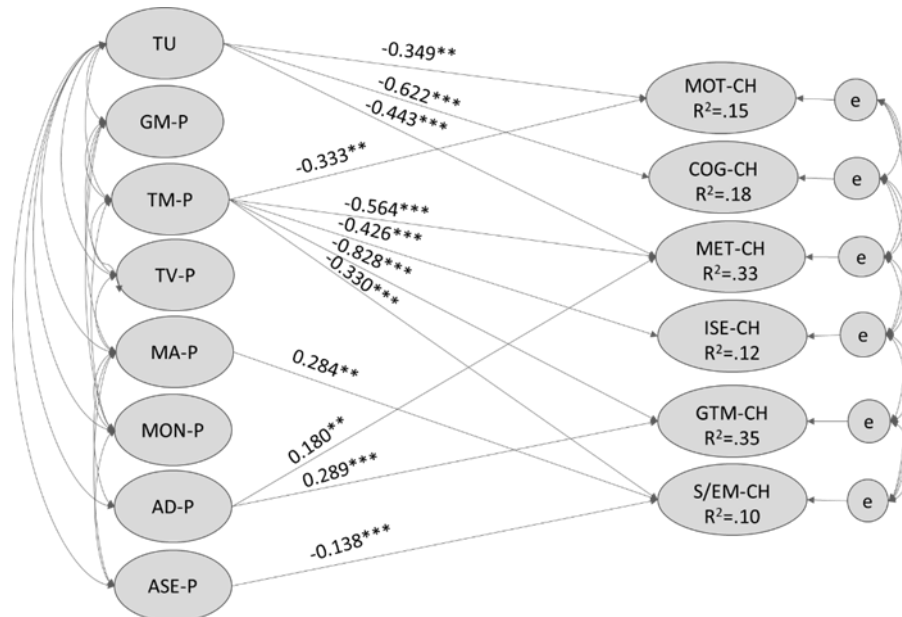
Abbreviations: Academic Challenges(-CH): (MOT:Motivation; COG:Cognitive; MET: Metacognitive; ISE: Initiating & Sustaining Engagement; GTM: Goal & Time Management; S/EM: Social and emotional).

Examining the Predictive Validity of the SRL Practices Scales for Academic Challenges

We conducted a second analysis to further test the predictive validity of the SRL practices scale for explaining variability in academic challenges. When academic challenges (SRL-C) are reported retrospectively after a study session, they can theoretically be modelled as the outcomes of poor SRL engagement (SRL-P). Therefore, academic challenges (SRL-C) were regressed on SRL practices (SRL-P) scales. We began with saturated models where all possible paths from predictors to the outcomes were included and then removed the insignificant paths to improve the model fit to the data. The pruned model was compared to the saturated model in terms of patterns of associations, global fit, and effect sizes to make sure removal of non-significant paths did not result in different results. Findings, presented in Figures 1, indicate that SRL practices were significantly associated with challenge outcomes. *Task understanding* negatively predicted *motivation*, *cognitive*, and *metacognitive* challenges. *Time management* practices negatively predicted all challenge areas except *cognitive* challenges, for which the effect was not statistically significant. *Academic social engagement* was negatively associated with *social and emotional* challenges. The negative associations obtained indicated that higher use of these self-regulatory practices was associated with lower reports of encountering academic challenges. On the contrary, *motivation appraisal* practices were found to positively associate with *social and emotional* challenges ($B=.28, p=.005$). Similarly, *adaptation* practices were found to positively associate with *metacognitive* ($B=.18, p=.003$) and *goal and time management* challenges ($B=.29, p<.001$).

Figure 1

SRL-Practices facets predicting Academic Challenges



Note. Model Fit Indices: $\chi^2(1787) = 2674.9^{***}$, CFI=.907, TLI=.902, RMSEA=.032 [.030, .035], SRMR= .055, estimator=MLR.

Correlation among predictors and residual correlations among outcome variables are presented in Table C4 in Appendix C.

Abbreviations: SRL-Practices (-P): (TU: Task Understanding; GM: Goal Management; ATV: Appraisals of Task Value; MA: Motivation Appraisal; TM: Time Management Practices; AD: Adaptation Practices; MON: Monitoring Practices; ASE: Academic Social Engagement).

Academic Challenges(-CH): (MOT: Motivation; COG: Cognitive; MET: Metacognitive; ISE: Initiating & Sustaining Engagement; GTM: Goal & Time Management; S/EM: Social and emotional).

To better understand the positive effect of *motivation appraisal* practices on *social and emotional* challenges, we removed the other two significant predictors of social and emotional challenges (i.e., TM & ASE) from the model. *Motivation appraisal* practices were no longer a significant predictor of *social and emotional* challenges in this new model, where it was the only predictor of GPA ($B = -.06, p = .50$). This observation indicated that when the effect of *time-management* and *academic social engagement* practices on *social and emotional* challenges is controlled, deploying higher *motivation appraisal* practices was positively associated with *social and emotional* challenges.

Similarly, after removing the other significant predictor of *goal and time management* challenges (i.e., *time management* practices) from the model, *adaptation* alone was no longer a significant predictor of *goal and time management* challenges ($B = 0.043, p = .52$). This means that only when the effect of *time-management* practices on *goal and time management* challenges is partialled out, deploying higher *adaptation* practices was associated with higher rates of perceived *goal and time management* challenges.

The same strategy was used for understanding the positive association observed between *adaptation* practices and *metacognitive* challenges. Again, *adaptation* was not a significant predictor of *metacognitive* challenge when it was the only predictor of *metacognitive* challenges ($B = -.09, p = .132$). This means that only when the effect of practices such as *time management* and *task understanding* on *metacognitive* challenge is controlled, enacting higher *adaptation* practices was associated with higher rates of *metacognitive* challenges. As shown in Figure 1, R^2 values depicting the proportion of variation in academic challenges which was explained by SRL practices ranges from .10 for the *social and emotional* challenges to .35 for *goal and time management* challenges.

Discussion

We judge the SRL-SPD scales are psychometrically adequate representations of students' perceptions about SRL-related constructs the scales are intended to represent. Specifically, a 4-factor solution depicting planning, metacognition and beliefs, foundational academic behaviours, and academic social engagement fit both the SRL-importance and SRL Self-efficacy scales. A 6-factor solution comprised of *metacognitive*, *cognitive*, *goal and time management*, *motivation*, *social and emotional challenges* and *initiating and sustaining engagement* challenges had an adequate fit for the SRL challenges scale. Finally, the SRL practices scale identified eight

distinct SRL practices with adequate fit including: *Task Understanding, Goal Management, Monitoring, Adapting, Time Management, Motivation Appraisal, Appraisals of Task Value, and Academic Social Engagement.*

Overall, three main findings demonstrate why the SRL-PSD scale exposes nuanced relations between SRL and academic performance. First, SRL practices, or what students do during studying, tended to be positively related to academic performance and negatively related to academic challenges reported during studying. Second, academic challenges were negatively related to academic performance. Together these findings suggest the predictive power of the SRL-PSD may differ depending on the specific SRL practice or academic challenge under examination.

While academic challenges consistently negatively predicted academic achievement (GPA), the relationship between SRL practices and academic achievement was not straightforward. Only four SRL practices (i.e., *task understanding, goal management, monitoring, and time management practices*) were positively associated with academic performance. Importantly, SRL practices that were not associated with academic performance (e.g., *adaptation, & academic social engagement*) were useful for predicting academic challenges students encountered during studying. Together these findings highlight that understanding the academic challenges students encounter during studying might provide context for explaining the impact of SRL practices on academic performance. Self-reports of study strategies or traces of studying events and event-sequences alone may provide an incomplete picture of self-regulated learning as practices situated in person-level beliefs and challenge experiences. SRL self-report measures, such as the MSLQ or LASSI capture students' perceptions of what they do without aligning those practices to beliefs or situational context that

warrant them. Findings are consistent with prior studies indicating that academic challenges may mediate the relationship between some SRL practices and academic performance (Bahena Olivares et al., accepted), as well as the relationship between Covid distress and academic performance (Hadwin et al., 2022).

Interestingly, adaptation practices were not significantly associated with GPA but were positively associated with academic challenges (goal and time management, and metacognitive challenges). This may seem surprising given that adaptation is the hallmark of SRL theory that posits successful learners adjust strategies and processes as the need arises. We posit that adaptation may only be productive for learning when it is strategically deployed in concert with other salient regulatory practices. Adapting or making changes to tactics is only indicative of strategic large-scale adaptation when it is coupled with metacognitive planning or feedback generated by monitoring and evaluation. Indeed, those who have worked with first year students have witnessed some students waste a lot of energy switching tactics/techniques and approaches without a strategic plan to guide those changes. Future research should examine the value of adaptation practices in the context of other metacognitive practices such as planning and monitoring (Berger & Karabenick, 2016). We posit that adaptation should not be used as an independent predictor but be modelled alongside other regulatory practices (e.g., task understanding, monitoring) that contextualize adaptation as part of a wholistic metacognitive response to learning situations.

Findings in the student success literature consistently report the importance of social and emotional experiences in learning and performance (Wong et al., 2024). An interesting finding regarding social and emotional aspects of studying emerged in this validation study was that Social and emotional challenges positively predicted academic performance only after

accounting for the effect of all other challenges on GPA. While it may seem counter-intuitive that more socio-emotional challenges were associated with higher academic performance, data for this study were collected in the context of a pandemic lockdown when most students were learning from home with limited social interactions with peers and instructors. During this time students may have been struggling with social and emotional challenges but simultaneously had nothing to do but study. This may have affected the association between social and emotional challenges and performance. In addition to cross validating the findings from this study, future research should examine the mechanisms through which social and emotional experiences together with other SRL beliefs and challenges contribute to academic performance.

A promising finding from this study regards the importance of planning for academic performance. Prioritizing planning (SRL-I: Planning), feeling confident about planning (SRL-SE: Planning), and engaging planning practices (*task understanding* and *goal management* practices) were all associated with academic performance. This finding extends past research indicating that planning practices such as task understanding contribute higher academic performance and lower levels of academic challenges (Bahena Olivares et al., accepted), higher task performance (Hadwin et al., 2001), and more effective teamwork (Bakhtiar et al., 2017). Future research should leverage the SRL-PSD scales to examine how planning beliefs and practices interact with challenges students encounter during studying to predict performance.

Finally, while past research indicates foundational academic behaviours (e.g., getting to class on time, engaging in class activities and discussions, etc.) are important for academic performance (Wong et al., 2024), this study is the first to demonstrate that beliefs about the foundational academic behaviours contribute to academic performance. Specifically, findings from this SRL-PSD study indicate that prioritizing foundational academic behaviours (SRL

importance) and feeling confident about those behaviours (SRL self-efficacy) are important for academic performance when the effects of other importance and self-efficacy beliefs were controlled. Considering the potency of these SRL-belief systems in relation to academic engagement and other SRL practices should be a focus of future research. It may be that foundational academic behaviours and beliefs about those behaviours play a more significant role in successfully self-regulating learning, than contemporary motivational research modelling has revealed.

Significance and Future Directions

We suggest six avenues for advancing research using the SRL-PSD. First, the SRL-PSD should be extended, refined, and validated across larger and more diverse samples. We expect this may lead to refining and augmenting the scale to include items that are suitable for more diverse samples.

Second, SRL-PSD data could be triangulated with multimodal data sources. Learning analytics are increasingly leveraging multimodal data for modeling learning processes, and self-regulated learning (Azevedo & Gašević, 2019). However, these advancements have not been coupled with advancements in self-reports of SRL. Rather than focusing on specific study skills or techniques, the SRL-PSD scale was intentionally designed to examine SRL beliefs, processes, and experiences. Triangulating these self-reports about studying with multimodal data sources and analytics has the potential to advance understandings about the alignment between what students do and what they intend to do when they manage learning in demanding situations. Additionally, learning analytics research increasingly recognizes the importance of cognitive, help-seeking, and foundational academic engagement, practices in predicting academic success

(Edwards et al., 2020). The role of these practices should be examined with greater detail in relation to the SRL practices currently represented in the SRL-PSD.

Third, it is important for research to examine the dynamic interplay between SRL beliefs and practices in predicting academic challenges and performance outcomes. For example, Bahena Olivares et al. (2024) examined the role task understanding practices play in predicting academic challenges and performance. Recognizing the interplay between phases and facets of self-regulated learning is important for testing models that explain performance outcomes. Given the tight theoretical coupling of task understanding practices and motivation appraisal and task value practices, research should examine ways these practices together contribute to academic challenges and performance outcomes.

Fourth, SRL is theorized to be a cyclical and recursive process wherein products of each cycle of SRL become conditions for deploying self-regulated learning practices in the future (e.g., Winne & Hadwin, 1998). From this perspective timing of data collection matters. For this validation study, data about SRL practices and challenges were collected simultaneously. Challenges were thereby treated as outcomes representing current challenge experiences or products of successful or failed SRL practices. However theoretically, challenges also serve as triggers for self-regulatory action (Hadwin et al., 2018; 2019; under review). Examining the effectiveness of specific SRL practices used to respond to academic challenges is an important avenue for future research. We are currently conducting daily diary research that examines (a) contingencies between the challenges students encounter during studying the specific practices they deploy to remediate those challenges, as well as (b) the effectiveness of those practices for remediating challenges.

Fifth, the SRL-PSD scale provides a set of useful measures for testing the effectiveness of SRL-based interventions for remediating the impact of academic challenges on achievement and retention outcomes (e.g., Hadwin et al., 2022). However, it may also be useful for examining academic outcomes associated with interventions designed to promote academic well-being and engagement. For example, we are leveraging the SRL-PSD subscales in an ongoing randomized control trial designed to deliver just in-time interventions to students at risk for substance misuse (UniVenture; Stewart et al., 2018-2025). The SRL-PSD offers opportunities to examine more nuanced impact of such an intervention on academic practices, beliefs and behaviours that may in turn remediate academic challenges associated with poor academic performance.

Future Considerations and Limitations

We strive to decolonize educational psychology research by engaging in reciprocal research practices that offer immediate relevance and reciprocity for participants. SRL-PSD scales have been designed to innovatively debrief participants by generating personalized visual summaries as immediate feedback including: (a) an individualized report summarizing subscale scores, (b) a general description of what those subscale scores are theoretically posited to reveal about learning strengths and weaknesses, (c) access to a debriefing session to support interpreting and critiquing personalized reports, (d) links to a strategy library with targeted strategies for specific challenges and SRL practice areas (see supplementary materials). While the diagnostic potential of the SRL-PSD was not empirically examined in this study, future research should compare the effectiveness of administering the SRL-PSD with and without self-diagnostic reports for learners. Research is needed to examine the impact of providing this type of self-diagnostic information to learners in terms of (a) increasing metacognitive awareness of SRL beliefs, practices and challenges faced during studying, and (b) stimulating selection and

deployment of strategies and techniques for remediating those challenges. The use of these individual reports for self-diagnosing and selecting strategies is beyond the scope of this study.

The SRL-PSD scale has been under development for several years. A preliminary iteration of the SRL practices subscale was tested by McCardle et al. (2015). At that time items were pruned to improve psychometric properties of the scale and to reduce the number of items. This approach balanced the need for a broad representation of SRL with the field-based demands of having an instrument that could be used in educational settings for diagnostic purposes. Our analyses indicate that the reliability of a couple of factors (e.g., *motivation appraisal* practices & *motivation* challenges) could be strengthened by adding one or two items back in. We acknowledge that response bias can be reduced by introducing reverse coded items, but the nature of this scales may limit the practicality of introducing reverse coded items. Preliminary attempts to add reverse-coded items to the SRL-C produced confusing items, leading to respondent error, so they were intentionally avoided during development of that scale.

To generate a comprehensive SRL instrument examining beliefs, practices and challenges that could also be used in practice meant compromises needed to be made on the number of items per scale. We had to balance the need for a comprehensive set of items with maintaining a questionnaire length that was reasonable for respondents to complete without response fatigue. Unfortunately, this led to using some double-barreled items in the SRL beliefs scales such as "Feeling like I belong to my program and university" and "Changing my beliefs or approaches as needed." For such items, we were interested in the overall importance and competency beliefs about belongingness and adaptation during academic studying. We acknowledge that to develop a practical scale, double-barreled items introduced other measurement errors. To avoid this issue, respondents could be asked to make three rating for each of the 31 practice items including: (a)

How important were these practices during studying, (b) How confident are you that you can successfully engage each of these practices, and (c) To what extent did you engage these practices.

The global fit of the measures was acceptable in this initial validation study. Further scale refinements to enhance the validity and reliability of the scales are needed. These improvements include both scale refinement and the use of advanced measurement models such as higher-order and bifactor CFA to model overarching factors underlying responses, along with their Exploratory Structural Equation Modeling (ESEM) equivalents to allow for the modeling of cross-loadings. This work is currently underway using a larger sample size.

The SRL-PSD was validated using a sample of undergraduate students, primarily first year students, recruited from the undergraduate psychology SONA pool, an educational psychology course focused on learning, motivation, and self-regulated learning, and supported by a learning strategy library available to all undergraduate students at the university. The sample includes students enrolled in a wide array of disciplines and degree programs across the university because first year psychology courses are recommended or required across programs. Hadwin et al. (2022) conducted a comparison between students who did and did not receive explicit training about self-regulated learning and found some notable differences. However, the SRL-PSD should be validated with a wider sample of students across institutions and contexts (e.g., Wu, 2021). Currently, data are being collected from undergraduate students across six universities, and four provinces in Canada.

We acknowledge issues using GPA as an aggregate measure of performance. GPA is the single most cited measure of academic performance used by institutions, admissions to advanced study and throughout the student success empirical literature. Our goal was to validate an

instrument that could be to measure factors contributing to student success at large. A focus of future research will be to examine variability across disciplines and campuses.

Importantly, data used to validate this instrument were collected during a worldwide pandemic which created extraordinary circumstances. Many students were engaged in university studies fully online for the first time in their academic lives. While we intuit this may have heightened experiences of academic challenges, it is also possible that it affected the factor structure of the items within scales. For example, the factor labelled social and emotional engagement in the final measurement model for the SRL practices scale (see Table 4) is an unexpected mix of items originally intended to tap into online learning engagement and behaviours. Future research should flesh out those behaviours more fully including items that are indicative of academic engagement across modes of delivery.

Finally, various cognitive strategies and practices were not included in the SRL practices scale. This was intentional because cognitive strategy items included in other instruments (e.g., LASSI, or MSLQ) are not universally relevant for a diverse array of academic tasks (Hadwin et al., 2001). Rather than focusing on specific study skills or techniques, the SRL-PSD scale was intentionally focused on examining complex SRL practices that could be appropriately deployed to strategically self-regulate learning across a wide range of tasks and discipline contexts. A paucity of scales exist that examine the cognitive processes (e.g., selecting for importance, elaborating, activating prior knowledge, organizing, and structuring concepts in meaningful ways) learners deploy to engage elaborate processing and recall, a limitation we identified many years ago (Hadwin & Winne, 1996). In response, we are currently piloting a separate scale designed specifically to target key cognitive processes comprising deeper processing without

conflating those practices with a limited set of techniques (e.g., highlighting, KWL maps, matrix notes) used to deploy those practices.

Conclusion

Multi-modal research about self-regulated learning (SRL) has accelerated over the past few years. However, the development of self-report methods has not kept pace with other methodological advancements. Accessing learners' subjective experiences uniquely contributes to interpreting multi-modal traces of actions and calibrating those traces with metacognitive knowledge used to guide strategy deployment (Craig et al., 2020). Classic self-report measures (e.g., LASSI, Weinstein et al. 1987; MSLQ, Pintrich et al. 1993) have limited utility for validating and testing the calibration of observed-multi-modal indicators because they (a) fail to collect data about why students choose to deploy specific practices during studying (Hadwin et al., 2001; McCardle & Hadwin, 2015, 2020), (b) neglect to examine metacognitive beliefs and practices used for strategic decision making and adaptation during studying (McCardle & Hadwin, 2015; Tock & Moxley, 2017), and (c) fail to provide a pragmatic framework for providing timely diagnostic feedback to learners and instructors. This study established the psychometric adequacy and predictive validity of the *Self-regulated Learning Profile and Self-diagnostic (SRL-PSD) instrument* with a sample of post-secondary students. The SRL-PSD instrument advances self-report measures of SRL by incorporating four independent scales (a) SRL self-efficacy, (b) SRL importance, (c) challenges experienced during studying, and (d) SRL practices underrepresented in past measures of SRL.

References

- Acee, T. W., Cho, Y., Kim, J.-I., & Weinstein, C. E. (2012). Relationships among properties of college students' self-set academic goals and academic achievement. *Educational Psychology*, 32(6), 681–698. <https://doi.org/10.1080/01443410.2012.712795>
- Adams, R. V., & Blair, E. (2019). Impact of time management behaviors on undergraduate engineering students' performance. *SAGE Open*, 9(1), 215824401882450. <https://doi.org/10.1177/2158244018824506>
- Ashton, K., Hoff, M. A., Lin, S., & Gonzales, C. (2021). LASSI's great adventure: A meta-analysis of the learning and study strategies inventory and academic outcomes. *Educational Research Review*, 34, 100407. <https://doi.org/10.1016/j.edurev.2021.100407>
- Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, 96, 207–210. <https://doi.org/10.1016/j.chb.2019.03.025>
- Bahena-Olivares, L. M., Rostampour, R., & Hadwin, A. F. (2024). Modelling the unique contributions of task understanding to academic challenges and performance [Manuscript accepted for publication].
- Ball, I., Banerjee, M., Holliman, A., & Tyndall, I. (2024). Investigating success in the transition to university: a systematic review of personal risk and protective factors influencing academic achievement. *Educational Psychology Review*, 36(2), 1-30. <https://doi.org/10.1007/s10648-024-09891-0>
- Bandura, A. (1997). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted Impact of Self-Efficacy Beliefs on Academic Functioning. *Child Development*, 67(3), 1206–1222. <https://doi.org/10.1111/j.1467-8624.1996.tb01791.x>

- Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-Efficacy Beliefs as Shapers of Children's Aspirations and Career Trajectories. *Child Development*, 72(1), 187–206. <https://doi.org/10.1111/1467-8624.00273>
- Bakhtiar, A., Webster, E.A. & Hadwin, A.F. Regulation and socio-emotional interactions in a positive and a negative group climate. *Metacognition Learning* 13, 57–90 (2018). <https://doi.org/10.1007/s11409-017-9178-x>
- Baumeister, R. F., & Leary, M. R. (2017). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Interpersonal development*, 57-89. <https://doi.org/10.1037/0033-2909.117.3.497>
- Beckman, K., Apps, T., Bennett, S., Dalgarno, B., Kennedy, G., & Lockyer, L. (2021). Self-regulation in open-ended online assignment tasks: The importance of initial task interpretation and goal setting. *Studies in Higher Education*, 46(4), 821-835. <https://doi.org/10.1080/03075079.2019.1654450>
- Berger, J. L., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational assessment*, 21(1), 19-33. <https://doi.org/10.1080/10627197.2015.1127751>
- Bernacki, M. L. (2018). Examining the cyclical, loosely sequenced, and contingent features of self-regulated learning: Trace data and their analysis. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 370–387). Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9781315697048-24>
- Biswas G., Baker, R.S., & Paquette, L. (2018). Datamining methods for assessing self-regulated learning. In D. H. Schunk & J. A. Greene (Eds) *Handbook of Self-Regulated Learning and Performance* (2nd ed). (pp. 388-404). Routledge. <https://doi.org/10.4324/9781315697048-25>
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied psychology*, 54(2), 199-231. <https://doi.org/10.1111/J.1464-0597.2005.00205.X>

- Bouchkioua, Z. (2021). Raising student's self-efficacy: the role of metacognitive training. *Psychology*, 12(11), 1900-1910. <https://doi.org/10.4236/psych.2021.1211115>
- Bowden, J. L. H., Tickle, L., & Naumann, K. (2021). The four pillars of tertiary student engagement and success: a holistic measurement approach. *Studies in Higher Education*, 46(6), 1207-1224. <https://doi.org/10.1080/03075079.2019.1672647>
- Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27, 1–13. <https://doi.org/10.1016/j.iheduc.2015.04.007>
- Butler, D. L. (2002). Qualitative approaches to investigating self-regulated learning: contributions and challenges. *Educational Psychologist*, 37(1), 59–63. https://doi.org/10.1207/S15326985EP3701_7
- Butler, D., & Winne P.H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245-281. <https://doi.org/10.3102/00346543065003245>
- Byrne, B. M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International journal of testing*, 1(1), 55-86. https://doi.org/10.1207/S15327574IJT0101_4
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Routledge.
- Callan, G. L., & Cleary, T. J. (2019). Examining cyclical phase relations and predictive influences of self-regulated learning processes on mathematics task performance. *Metacognition and Learning*, 14(1), 43-63. <https://doi.org/10.1007/s11409-019-09191-x>
- Colthorpe, K., Sharifirad, T., Ainscough, L., Anderson, S., & Zimbardi, K. (2018). Prompting undergraduate students' metacognition of learning: Implementing 'meta-learning'

- assessment tasks in the biomedical sciences. *Assessment and Evaluation in Higher Education*, 43(2), 272-285. <https://doi.org/10.1080/02602938.2017.1334872>
- Conti, R. (2000). College goals: Do self-determined and carefully considered goals predict intrinsic motivation, academic performance, and adjustment during the first semester? *Social Psychology of Education*, 4(2), 189-211. <https://doi.org/10.1023/A:1009607907509>
- Corno, L. (1986). The metacognitive control components of self-regulated learning. *Contemporary Educational Psychology*, 11(4), 333-346. <https://doi.org/10.1016/0361-476X%2886%2990029-9>
- Cosnefroy, L., Fenouillet, F., Mazé, C., & Bonnefoy, B. (2018). On the relationship between the forethought phase of self-regulated learning and self-regulation failure. *Issues in Educational Research*, 28(2), 329.
- Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning*, 15(2), 155-213. <https://doi.org/10.1007/s11409-020-09222-y>
- de Bruijn-Smolers, M., Timmers, C. F., Gawke, J. C. L., Schoonman, W., & Born, M. P. (2016). Effective self-regulatory processes in higher education: Research findings and future directions. A systematic review. *Studies in Higher Education*, 41, 139–158. <https://doi.org/10.1080/03075079.2014.915302>
- Dinsmore, D. L. (2017). Toward a dynamic, multidimensional research framework for strategic processing. *Educational Psychology Review*, 29(2), 235-268. <https://doi.org/10.1007/s10648-017-9407-5>
- Dixson, M. D. (2015). Measuring student engagement in the online course: The Online Student Engagement scale (OSE). *Online Learning*, 19(4), n4.
- Durik, A. M., Lovejoy, C. M., & Johnson, S. J. (2009). A longitudinal study of achievement goals for college in general: Predicting cumulative GPA and diversity in course selection.

- Contemporary Educational Psychology*, 34(2), 113–119.
<https://doi.org/10.1016/j.cedpsych.2008.11.002> Fong, C. J., Krou, M. R., Johnston.
<https://doi.org/10.1016/J.CEDPSYCH.2008.11.002>
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1), 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Edwards, R., Davis, S., Hadwin, A. F., & Milford, T. M. (2020). Exploring student engagement factors in a blended undergraduate course. *The Canadian Journal for the Scholarship of Teaching and Learning*, 11(3), 1-22. <https://doi.org/10.5206/cjsotl-rcacea.2020.3.8293>
- Finney, S. J., & DiStefano, C. (2008). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. D. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 269-314). Information Age Publishing.
- Fokkens-Bruinsma, M., Vermue, C., Deinum, J. F., & Van Rooij, E. (2021). First-year academic achievement: the role of academic self-efficacy, self-regulated learning and beyond classroom engagement. *Assessment & Evaluation in Higher Education*, 46(7), 1115-1126. <https://doi.org/10.1080/02602938.2020.1845606>
- Fong, C. J., Krou, M. R., Johnston-Ashton, K., Hoff, M. A., Lin, S., & Gonzales, C. (2021). LASSI's great adventure: A meta-analysis of the learning and study strategies inventory and academic outcomes. *Educational Research Review*, 34, 100407.
<https://doi.org/10.1016/j.edurev.2021.100407>
- Greene, J. A., Hutchison, L. A., Costa, L., & Crompton, H. (2012). Investigating how college students' task definitions and plans relate to self-regulated learning processing and understanding of a complex science topic. *Contemporary Educational Psychology*, 37(4), 307-320. <https://doi.org/10.1016/j.cedpsych.2012.02.002>
- Hadwin, A. F. (2021). Commentary and future directions: What can multi-modal data reveal about temporal and adaptive processes in self-regulated learning? *Learning and Instruction*, 72, 101287. <https://doi.org/10.1016/j.learninstruc.2019.101287>

- Hadwin, A. F., Bakhtiar, A., & Davis, S. K. (2018, April). Promoting & researching SRL: Strategic and adaptive responses to new situations and challenges. In M. Bernacki (Chair), *An integrative and comparative analysis of approaches to developing undergraduates learning skills*. Symposium conducted at the meeting of the American Educational Research Association, New York, NY.
- Hadwin, A. F., Bakhtiar, A., & Miller, M. (2018). Challenges in online collaboration: Effects of scripting shared task perceptions. *International Journal of Computer-Supported Collaborative Learning*, 13(3), 301-329. <https://doi.org/10.1007/s11412-018-9279-9>
- Hadwin, A. F., Bakhtiar, A., Davis, S., & Greco, S. (2019a, June). *Promoting and researching self-regulation of learning as strategic and adaptive responses to new situations and challenges* [Paper presentation]. Adaptive and Maladaptive Self-regulation Symposium, Annual Meeting of the Canadian Society for the Study of Education, Vancouver, BC, Canada.
- Hadwin, A. F., Davis, S. K., Bakhtiar, A., & Winne, P. H. (2019b). Academic challenges as opportunities to learn to self-regulate learning. In H. Askell-Williams & J. Orrell (Eds.), *Problem solving for teaching and learning: A festschrift for Emeritus Professor Mike Lawson* (pp. 34-48). Routledge. <https://doi.org/10.4324/9780429400902>
- Hadwin, A. F., Oshige, M., Miller, M., Fior, M. N., & Tupper, K. (2008). *Examining the agreement between student and instructor task perceptions in a complex engineering design task* [Poster presentation]. American Education Research Association, New York, NY.
- Hadwin, A. F., Oshige, M., Miller, M., & Wild, P. M. (2009). Examining the agreement between student and instructor task perceptions in a complex engineering design task. *Proceedings of the Canadian Engineering Education Association (CEEAA)*.
- Hadwin, A. F., Sukhawathanakul, P., Rostampour, R., & Bahena Olivares, M. (2022). Do self-regulated learning practices and intervention mitigate the impact of academic challenges

- and Covid distress on academic performance during online learning? *Frontiers in Psychology*, 13, Article 813529. <https://doi.org/10.3389/fpsyg.2022.813529>
- Hadwin, A. F., & Winne, P. H. (1996). Study strategies have meager support: A review with recommendations for implementation. *The Journal of Higher Education*, 67(6), 692-715. <https://doi.org/10.1080/00221546.1996.11774821>
- Hadwin, A. F., & Winne, P. H. (2012). Promoting learning skills in undergraduate students. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 201-227). Cambridge University Press. <https://doi.org/10.1017/CBO9781139048224.013>
- Hadwin, A. F., Winne, P. H., Stockley, D. B., Nesbit, J. C., & Woszczyna, C. (2001). Context moderates students' self-reports about how they study. *Journal of Educational Psychology*, 93(3), 477-487. <https://doi.org/10.1037/0022-0663.93.3.477>
- Hair, J., Black, W. C., Babin, B. J. & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, New Jersey: Pearson Educational International.
- Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A Measure of College Student Course Engagement. *The Journal of Educational Research*, 98(3), 184-192. <https://doi.org/10.3200/JOER.98.3.184-192>
- Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But.... *Communication Methods and Measures*, 14(1), 1-24. <https://doi.org/10.1080/19312458.2020.1718629>
- Hayes, A. F., & Rockwood, N. J. (2020). Conditional process analysis: Concepts, computation, and advances in the modeling of the contingencies of mechanisms. *American Behavioral Scientist*, 64(1), 19-54. <https://doi.org/10.1177/0002764219859633>
- Her, P., & Thompson, M. N. (2022). Examining the persistence intentions of college students of color. *Journal of Career Assessment*, 30(1), 41-59. <https://doi.org/10.1177/10690727211010382>

- Howell, A. J., & Buro, K. (2009). Implicit beliefs, achievement goals, and procrastination: A mediational analysis. *Learning and Individual Differences*, 19(1), 151–154.
<https://doi.org/10.1016/j.lindif.2008.08.006>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2017). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*, 109(3), 387-404. <https://doi.org/10.1037/edu0000146>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer. <https://doi.org/10.1007/978-1-4614-7138-7>
- Jansen, R. S., van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (2019). Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. *Educational Research Review*, 28, 100292.
<https://doi.org/10.1016/j.edurev.2019.100292>
- Järvelä, S., & Bannert, M. (2021). Temporal and adaptive processes of regulated learning - what can multimodal data tell? *Learning and Instruction*, 72, 101268. <https://doi.org/10.1016/j.learninstruc.2019.101268>
- Järvelä, S., Järvenoja, H., & Malmberg, J. (2019). Capturing the dynamic and cyclical nature of regulation: Methodological Progress in understanding socially shared regulation in learning. *International Journal of Computer-Supported Collaborative Learning*, 14 (4), 425-441. <https://doi.org/10.1007/s11412-019-09313-2>
- Jo, E., Lee, M., & Lee, W. (2021). Influence of Early Adolescents' Mastery Goals on Inter- and Intra-Individual Relationships Between Metacognitive Monitoring and Academic Outcomes. *The Journal of Early Adolescence*, 41(8), 1203–1227.
<https://doi.org/10.1177/0272431621989815>

- Keyes C. L. M. (2005). Mental illness and/or mental health? Investigating the axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, 73, 539–548. <https://doi.org/10.1037/0022-006X.73.3.539>
- Kim, Y., Brady, A. C., & Wolters, C. A. (2020). College students' regulation of cognition, motivation, behavior, and context: Distinct or overlapping processes? *Learning and Individual Differences*, 80, 101872. <https://doi.org/10.1016/j.lindif.2020.101872>
- Klassen, R. M. (2010). Confidence to manage learning: The self-efficacy for self-regulated learning of early adolescents with learning disabilities. *Learning Disability Quarterly*, 33(1), 19-30. <https://doi.org/10.1177/073194871003300102>
- Klassen, R. M., Ang, R. P., Chong, W. H., Krawchuk, L. L., Huan, V. S., Wong, I. Y. F., & Yeo, L. S. (2010). Academic procrastination in two settings: Motivation correlates, behavioral patterns, and negative impact of procrastination in Canada and Singapore. *Applied Psychology*, 59(3), 361-379. <https://doi.org/10.1111/j.1464-0597.2009.00394.x>
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. (4th ed.). Guilford Press.
- Knoesen, R., & Naudé, L. (2017). Experiences of flourishing and languishing during the first year at university. *Journal of Mental Health*, 27(3), 269–278. <https://doi.org/10.1080/09638237.2017.1370635>
- Koivuniemi, M., Panadero, E., Malmberg, J., & Järvelä, S. (2017). Higher education students' learning challenges and regulatory skills in different learning situations / desafíos de aprendizaje y habilidades de regulación en distintas situaciones de aprendizaje en estudiantes de educación superior. *Infancia y Aprendizaje*, 40(1), 19-55. <https://doi.org/10.1080/02103702.2016.1272874>
- Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New directions for institutional research*, 141, 5-20. <https://doi.org/10.1002/ir.283>

- Lance, C.E., Butts, M.M. & Michels, L.C., 2006. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods*, 9(2), pp.202–220. <https://doi.org/10.1177/1094428105284919>
- Limesurvey GmbH. / LimeSurvey: An Open-Source survey tool /LimeSurvey GmbH, Hamburg, Germany. URL <http://www.limesurvey.org>
- Marjon Fokkens-Bruinsma, Carlien Vermue, Jan-Folkert Deinum & Els van Rooij (2021) First-year academic achievement: the role of academic self-efficacy, self-regulated learning and beyond classroom engagement, *Assessment & Evaluation in Higher Education*, 46:7, 1115-1126, DOI: [10.1080/02602938.2020.1845606](https://doi.org/10.1080/02602938.2020.1845606)
- McCardle, L., & Hadwin, A. F. (2015). Using multiple, contextualized data sources to measure learners' perceptions of their self-regulated learning. *Metacognition and Learning*, 10(1), 43-75. <https://doi.org/10.1007/S11409-014-9132-0>
- McCardle, L., Webster, E.A., Haffey, A. & Hadwin, A.F. (2017). Examining students' self-set goals for self-regulated learning: Goal properties and patterns. *Studies in Higher Education*, 42(11), 2153-2169. <https://doi.org/10.1080/03075079.2015.1135117>
- Miller, M. F. W. (2009). *Predicting university students' performance on a complex task: Does task understanding moderate the influence of self-efficacy?* Unpublished Master's thesis. University of Victoria, Victoria, British Columbia, Canada.
- Oshige, M., Hadwin, A. F., Fior, M., Tupper, K., Miller, M. (2007). Why do students fail to understand the task assigned? Examining how attuned students are with a teacher across three levels of task-understanding. [Paper presentation]. Conference for the Canadian Society for the Study of Education.
- Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74-98. <https://doi.org/10.1016/j.edurev.2017.08.004>
- Panadero, E., Klug, J., & Järvelä, S. (2016) Third wave of measurement in the self-regulated learning field: when measurement and intervention come hand in hand, *Scandinavian*

- Journal of Educational Research*, 60 (6), 723-735.
<https://doi.org/10.1080/00313831.2015.1066436>
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, 53(3), 801-813.
<https://doi.org/10.1177/0013164493053003024>
- Putarek, V., & Pavlin-Bernardić, N. (2020). The role of self-efficacy for self-regulated learning, achievement goals, and engagement in academic cheating. *European Journal of Psychology of Education*, 35(3), 647-671. <https://doi.org/10.1007/s10212-019-00443-7>
- Roth, Ogrin, S., & Schmitz, B. (2015). Assessing self-regulated learning in higher education: a systematic literature review of self-report instruments. *Educational Assessment, Evaluation and Accountability*, 28(3), 225–250. <https://doi.org/10.1007/s11092-015-9229-2>
- Rosenzweig, E. Q., Wigfield, A., & Hulleman, C. S. (2020). More useful or not so bad? examining the effects of utility value and cost reduction interventions in college physics. *Journal of Educational Psychology*, 112(1), 166-182.
<https://doi.org/10.1037/edu0000370>
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2)). <https://doi.org/10.18637/jss.v048.i02>
- Roussel, P., Elliot, A. J., & Feltman, R. (2011). The influence of achievement goals and social goals on help-seeking from peers in an academic context. *Learning and Instruction*, 21(3), 394–402. <https://doi.org/10.1016/j.learninstruc.2010.05.003>
- Schellings, G., & Van Hout-Wolters, B. (2011). Measuring strategy use with self-report instruments: theoretical and empirical considerations. *Metacognition and Learning*, 6, 83-90. <https://doi.org/10.1007/s11409-011-9081-9>
- Schumacker, R. E., & Lomax, R. G. (2004). *A Beginner's guide to structural equation modeling* (2nd ed.). New York, NY: Psychology Press. <https://doi.org/10.4324/9781410610904>

- Schunk, D. H., & DiBenedetto, M. K. (2016). Self-efficacy theory in education. In K. R. Wentzel, & D. B. Miele (Eds.). *Handbook of motivation at school* (pp. 34–53). (2nd ed.). New York: Routledge. <https://doi.org/10.4324/9781315773384.CH3>
- Schunk, D. H., & Ertmer, P. A. (2000). Self-regulation and academic learning: Self-efficacy enhancing interventions. In *Handbook of self-regulation* (pp. 631-649). Academic Press.
- Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: what we know and where we need to go. *Psychological bulletin*, 137(3), 421. <https://doi.org/10.1037/a0022777>
- Stewart, S. H., Livet, A., Arjona, R. N., Keough, M. T., Thompson, K., Krank, M., ... & Conrod, P. J. (2021). Adapting the Preventure Substance Use Prevention Program into Online “Univenture”: Adaptation Process and Feasibility Findings. In *CCSA’s Issues of Substance 2021 conference*.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (5th ed.)*. Boston,. MA: Allyn & Bacon/Pearson Education.
- Thibodeaux, J., Deutsch, A., Kitsantas, A., & Winsler, A. (2017). First-year college students’ time use: Relations with self-regulation and GPA. *Journal of Advanced Academics*, 28(1), 5-27. <https://doi.org/10.1177/1932202X16676860>
- Tinto, V. (1982). Limits of theory and practice in student attrition. *The journal of higher education*, 53(6), 687-700. <https://doi.org/10.2307/1981525>
- Tock, J. L., & Moxley, J. H. (2017). A comprehensive reanalysis of the metacognitive self-regulation scale from the MSLQ. *Metacognition and Learning*, 12(1), 79 111. <https://doi.org/10.1007/s11409-016-9161-y>
- Trautner, M., & Schwinger, M. (2020). Integrating the concepts self-efficacy and motivation regulation: How do self-efficacy beliefs for motivation regulation influence self-regulatory success? *Learning and Individual Differences*, 80, 101890. <https://doi.org/10.1016/j.lindif.2020.101890>

- Usher, E. L., & Pajares, F. (2008). Self-efficacy for self-regulated learning: A validation study. *Educational and Psychological Measurement*, 68(3), 443–463.
<https://doi.org/10.1177/0013164407308475>
- Univenture (2018-2025). SSHRC Partnership Grant [Stewart, S. (PI); Conrad, P., Hadwin, A.F. Keough, M., Krank, M., Thompson, K. (CO-PI)]: *Univenture: A Partnership to Address Heavy Drinking and Other Substance Misuse on Canadian University Campuses*.
- Veldman, J., Meeussen, L., & van Laar, C. (2019). A social identity *perspective* on the social-class achievement gap: Academic and social adjustment in the transition to university. *Group Processes & Intergroup Relations*, 22(3), 403-418.
<https://doi.org/10.1177/1368430218813442>
- Weinstein, C. E., Schulte, A. C., & Palmer, D. R. (1987). Learning and study strategies inventory. Clearwater, FL: H & HPublishing.
- Wigfield A., Tonks, S. M., Klauda, S. L. (2016). Expectancy-Value Theory. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of Motivation at School* (2nd ed., pp. 55-74). Routledge.
<https://doi.org/10.4324/9781315773384>
- Winne, P. H. (2010). Improving Measurements of Self-Regulated Learning. *Educational Psychologist*, 45(4), 267–276. <https://doi.org/10.1080/00461520.2010.517150>
- Winne, P. H. (2018). Cognition and metacognition within self-regulated learning. In D. Schunk & J. Greene (Eds.), *Handbook of self-regulation of learning and performance*. (2nd ed., pp. 36-48). New York, NY: Routledge. <https://doi.org/10.4324/9781315697048.CH3>
- Winne, P. H. (2019). Paradigmatic dimensions of instrumentation and analytic methods in research on self-regulated learning. *Computers in Human Behavior*, 96, 285-290.
<https://doi.org/10.1016/j.chb.2019.03.026>
- Winne, P. H. (2020a). A proposed remedy for grievances about self-report methodologies. *Frontline Learning Research*, 8, 165-174. <https://doi.org/10.14786/flr.v8i3.625>
- Winne, P. H. (2020b). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, 112. <https://doi.org/10.1016/j.chb.2020.106457>

- Winne, P. H. (2022). Modeling self-regulated learning as learners doing learning science: How trace data and learning analytics help develop skills for self-regulated learning. *Metacognition and Learning*, 17(3), 773-791. <https://doi.org/10.1007/s11409-022-09305-y>
- Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 277–304). Mahwah, NJ: Erlbaum. <https://doi.org/10.4324/9781410602350-19>
- Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self-reports about study tactics and achievement. *Contemporary Educational Psychology*, 27(4), 551-572. [https://doi.org/10.1016/S0361-476X\(02\)00006-1](https://doi.org/10.1016/S0361-476X(02)00006-1)
- Winne, P. H., & Marzouk, Z. (2019). Learning strategies and self-regulated learning. In J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 696-715). New York, NY: Cambridge University Press. <https://doi.org/10.1017/9781108235631.028>
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In P. Pintrich, M. Boekaerts, & M. Zeidner (eds), *Handbook of self-regulation* (pp. 531-566). Academic Press. <https://doi.org/10.1016/B978-012109890-2%2F50045-7>
- Winne, P. H., Zhou, M., & Egan, R. (2011). Designing assessments of self-regulated learning. In G. Shraw & D.R. Robinson (Eds.), *Assessment of higher order thinking skills* (pp. 89–118). Charlotte: Information Age.
- Wolters, C. A. (1998). Self-regulated learning and college students' regulation of motivation. *Journal of Educational Psychology*, 90(2), 224–235. <https://doi.org/10.1037/0022-0663.90.2.224>
- Wolters, C. A., & Benzon, M. B. (2013). Assessing and Predicting College Students' Use of Strategies for the Self-Regulation of Motivation. *The Journal of Experimental Education*, 81(2), 199–221. <https://doi.org/10.1080/00220973.2012.699901>

- Wolters, C. A., Brady, A. C. (2020). College Students' Time Management: A Self-Regulated Learning Perspective. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-020-09519-z>
- Wolters, C. A. & Won, S. (2018). Validity and the use of self-report questionnaires to assess self-regulated learning. In D. H. Schunk & J. A. Greene (Eds.) *Handbook of Self-Regulated Learning and Performance*. (2nd ed.) (pp. 307-322). Routledge. <https://doi.org/10.4324/9781315697048.CH20>
- Won, S., Hensley, L. C., & Wolters, C. A. (2021). Brief research report: Sense of belonging and academic help-seeking as self-regulated learning. *The Journal of Experimental Education*, 89(1), 112-124. <https://doi.org/10.1080/00220973.2019.1703095>
- Won, S., & Yu, S. L. (2018). Relations of perceived parental autonomy support and control with adolescents' academic time management and procrastination. *Learning and Individual Differences*, 61, 205-215. <https://doi.org/10.1016/j.lindif.2017.12.001>
- Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of Educational Psychology*, 116(1), 48–75. <https://doi.org/10.1037/edu0000833>
- Wu M., (2021). *Culture and Self-regulated learning: Exploring cultural influences on Chinese international and Canadian domestic undergraduate students' engagement in self-regulated learning*. [Master's thesis, University of Victoria]. <http://hdl.handle.net/1828/12558>.
- Zhou, M., & Kam, C. C. S. (2017). Trait procrastination, self-efficacy and achievement goals: the mediation role of boredom coping strategies. *Educational Psychology*, 37(7), 854–872. <https://doi.org/10.1080/01443410.2017.1293801>
- Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, 11(4), 307-313. [https://doi.org/10.1016/0361-476X\(86\)90027-5](https://doi.org/10.1016/0361-476X(86)90027-5)

- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329-339. <https://doi.org/10.1037/0022-0663.81.3.329>
- Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational psychologist*, 25(1), 3-17. https://doi.org/10.1207/s15326985ep2501_2
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). San Diego, CA: Academic Press. <https://doi.org/10.1016/B978-012109890-2%2F50031-7>
- Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning-strategies. *American Educational Research Journal*, 23(4), 614-628. <https://doi.org/10.3102/00028312023004614>
- Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and an overview. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15-26). Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9780203839010-4>
- Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, 29(3), 663. <https://doi.org/10.2307/1163261>

Appendix A
SRL-PSD: Self-Regulated Learning Profiles and Self-Diagnostic Instrument

<i>SRL Importance Items.</i> <i>How important were these practices to you in the last two week</i> (Never Important, Rarely Important, Sometimes Important, Usually Important, Always Important)	
Planning	Thoroughly understanding what I need to do Managing my goals Managing my time
Metacognition & Beliefs	Changing my beliefs or approaches as needed Finding value or usefulness in the study session Thinking about my effort and feelings related to this study session. Monitoring my Learning
Foundational Academic Behaviours	Showing up for class or online meetings Getting myself to do school work on time Participating in learning activities or discussions
Academic Social Engagement	Having fun and enjoying my academic work Making friends and getting to know people Feeling like I belong to my program and university.
<i>SRL Self-Efficacy</i> <i>How confident are you that you can successfully engage each of these practices in upcoming study sessions in the next two weeks</i> (Not Confident, Slightly Confident, Very Confident, Completely Confident)	
Planning	Thoroughly understanding what I need to do Managing my goals Managing my time
Metacognition & Beliefs	Changing my beliefs or approaches as needed Finding value or usefulness in the study session Thinking about my effort and feelings related to this study session. Monitoring my Learning
Foundational Academic Behaviours	Showing up for class or online meetings Getting myself to do school work on time Participating in learning activities or discussions
Academic Social Engagement	Having fun and enjoying my academic work Making friends and getting to know people Feeling like I belong to my program and university.
<i>Self-regulated learning Practices (SRL-P)</i> <i>Please rate your level of agreement with each following statement.. Over the last two weeks, when I did academic work, I...</i> (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree)	
Goal Management	Set goals for my work Made goals to learn, understand, or remember Set goals that will be useful for checking my progress Identified specific content, ideas, or terms in my goals Assessed my goal attainment
Task Understanding	Asked myself if I know what is important to learn Made sure I understand terminology used in instructions Thought about the professor's expectations and standards Considered what knowledge or big ideas I should learn or demonstrate Thought about relevant information and resources
Appraisals of Task Value	Thought about why we are being asked to know this stuff Reflected on why this work is important Made judgments about the usefulness or value of the content
Motivation Appraisal	Assessed if think I can do it

	Assessed my feelings for the task Evaluated the effort I was putting in
Metacognitive Monitoring	Asked myself if I understand what I am supposed to be doing Asked myself if I was remembering Asked myself if I am understanding the material
Metacognitive Control & Adaptation	Changed my understanding of the task at hand Modified my plans for the task Switched to a different strategy or approach Changed my feelings about the task Altered the level of effort I put in Modified my beliefs about how well I would do my tasks
Time Management	Checked to see if I am staying on time Chose goals that can be completed in one reasonable work session Created a timeline or schedule
Academic Social Engagement	Got to know people in the class Had fun in online chats, discussions or via email with the instructor or other students Helped fellow students
SRL Challenges (SRP-C) Items The following are the list of things that can make studying and learning at university difficult. Think about the past 2 weeks when you rate your level of agreement with each following statement. Over the past two weeks, I struggled with... (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree)	
Metacognitive Challenges	Breaking tasks down Interpreting tasks and expectations Knowing how to tackle tasks Knowing how or when to fix strategies/study skills Prioritizing or figuring out what is important Monitoring my task or learning progress Setting goals for studying Applying the right strategies/techniques properly
Social & Emotional challenges	Feeling connected Avoiding feeling lonely Finding enjoyment at university Managing my emotions/feelings Managing relationships Taking care of my mental health and wellbeing
Cognitive Challenges	Applying my learning Elaborating and connecting what I learned Grappling with language and communication Identifying important ideas Remembering things I learned Understanding learning materials and concepts
Challenges Initiating-sustaining engagement	Wanting to do my schoolwork Avoiding putting things off (procrastinating) Focusing or concentrating Managing distractions
Goal and time Management Challenges	Keeping my commitments or goals Organizing time and tasks Managing my time, tasks, or goals Meeting deadlines
Motivation Challenges	Believing I can do my work Feeling like my work was worth doing Persisting when things got tough

Appendix B
Supplementary Statistics related to Measurement Models

Table B1*Abbreviation Used for Labling Scales*

Abbreviation	Full Name of the scale
SRL-I	Self-regulated Learning-Importance
-Plan	Planning
-MB	Metacognition and Beliefs
-FAB	Foundational Academic Behaviours
-ASE	Academic Social Engagement
SRL-SE	Self-regulated Learning-Self-Efficacy
-Plan	Planning
-MB	Metacognition and Beliefs
-FAB	Foundational Academic Behaviours
-ASE	Academic Social Engagement
SRL-P	Self-regulated Learning-Practices
-TU	Task Understanding
-GM	Goal Management
-TM	Time Management
-ATV	Appraisals of Task Value
-MA	Motivation Appraisals
-MON	Monitoring
-AD	Adaptation
-ASE	Academic Social Engagement
SRL-CH	Self-regulated Learning-Challenges
-MOT	Motivation
-COG	Cognitive
-MET	Metacognitive
-ISE	Initiating and Sustaining Engagement
-GTM	Goal and Time Management
-S/EM	Social and Emotional

Table B2
Descriptive Statistics

	Valid	Missing	Mean	Std. Deviation	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis	Minimum	Maximum
GPA	473	0	6.627	1.842	-0.848	0.112	0.143	0.224	1.000	9.000
SRL-I-Plan1	473	0	3.941	0.768	-0.491	0.112	0.223	0.224	1.000	5.000
SRL-I-Plan2	473	0	3.748	0.840	-0.467	0.112	0.272	0.224	1.000	5.000
SRL-I-Plan3	473	0	3.926	0.874	-0.621	0.112	0.092	0.224	1.000	5.000
SRL-I-MB1	473	0	2.983	0.930	0.081	0.112	-0.498	0.224	1.000	5.000
SRL-I-MB2	473	0	3.359	0.990	-0.191	0.112	-0.461	0.224	1.000	5.000
SRL-I-MB3	473	0	2.962	1.081	0.035	0.112	-0.644	0.224	1.000	5.000
SRL-I-MB4	473	0	3.182	0.981	-0.236	0.112	-0.520	0.224	1.000	5.000
SRL-I-FAB1	473	0	3.909	1.075	-0.806	0.112	-0.138	0.224	1.000	5.000
SRL-I-FAB2	473	0	4.296	0.862	-1.328	0.112	1.669	0.224	1.000	5.000
SRL-I-FAB3	473	0	3.205	1.084	-0.195	0.112	-0.522	0.224	1.000	5.000
SRL-I-ASE1	473	0	2.708	1.108	0.268	0.112	-0.555	0.224	1.000	5.000
SRL-I-ASE2	473	0	2.256	1.186	0.703	0.112	-0.456	0.224	1.000	5.000
SRL-I-ASE3	473	0	2.662	1.166	0.305	0.112	-0.771	0.224	1.000	5.000
SRL-SE-Plan1	473	0	3.397	0.882	-0.365	0.112	0.188	0.224	1.000	5.000
SRL-SE-Plan2	473	0	3.254	0.892	-0.125	0.112	-0.046	0.224	1.000	5.000
SRL-SE-Plan3	473	0	3.000	1.060	-0.075	0.112	-0.586	0.224	1.000	5.000
SRL-SE-MB1	473	0	2.905	0.903	-0.158	0.112	-0.270	0.224	1.000	5.000
SRL-SE-MB2	473	0	2.987	0.957	-0.048	0.112	-0.409	0.224	1.000	5.000
SRL-SE-MB3	473	0	2.829	0.995	-0.027	0.112	-0.614	0.224	1.000	5.000
SRL-SE-MB4	473	0	3.030	0.958	-0.103	0.112	-0.352	0.224	1.000	5.000
SRL-SE-FAB1	473	0	3.772	1.167	-0.727	0.112	-0.380	0.224	1.000	5.000
SRL-SE-FAB2	473	0	3.767	1.107	-0.678	0.112	-0.326	0.224	1.000	5.000
SRL-SE-FAB3	473	0	3.074	1.139	-0.007	0.112	-0.791	0.224	1.000	5.000
SRL-SE-ASE1	473	0	2.486	1.011	0.297	0.112	-0.430	0.224	1.000	5.000
SRL-SE-ASE2	473	0	2.258	1.203	0.712	0.112	-0.401	0.224	1.000	5.000
SRL-SE-ASE3	473	0	2.539	1.153	0.375	0.112	-0.620	0.224	1.000	5.000
SRL-P-TU1	473	0	3.674	0.913	-0.717	0.112	0.227	0.224	1.000	5.000
SRL-P-TU2	473	0	3.979	0.733	-0.809	0.112	1.439	0.224	1.000	5.000
SRL-P-TU3	473	0	4.032	0.816	-0.927	0.112	1.364	0.224	1.000	5.000
SRL-P-TU4	473	0	3.922	0.808	-0.752	0.112	0.915	0.224	1.000	5.000
SRL-P-TU5	473	0	3.772	0.875	-0.835	0.112	0.787	0.224	1.000	5.000
SRL-P-GM1	473	0	3.920	0.910	-0.823	0.112	0.575	0.224	1.000	5.000
SRL-P-GM2	473	0	3.569	0.996	-0.449	0.112	-0.362	0.224	1.000	5.000
SRL-P-GM3	473	0	3.275	1.091	-0.130	0.112	-0.884	0.224	1.000	5.000
SRL-P-GM4	473	0	3.391	0.971	-0.305	0.112	-0.633	0.224	1.000	5.000
SRL-P-GM5	473	0	3.425	1.025	-0.402	0.112	-0.563	0.224	1.000	5.000
SRL-P-ATV1	473	0	3.391	1.122	-0.435	0.112	-0.734	0.224	1.000	5.000
SRL-P-ATV2	473	0	3.484	0.959	-0.534	0.112	-0.369	0.224	1.000	5.000
SRL-P-ATV3	473	0	3.744	0.883	-0.849	0.112	0.607	0.224	1.000	5.000
SRL-P-TM1	473	0	3.911	0.928	-0.958	0.112	0.836	0.224	1.000	5.000
SRL-P-TM2	473	0	3.414	1.113	-0.427	0.112	-0.699	0.224	1.000	5.000
SRL-P-TM3	473	0	3.865	1.164	-0.917	0.112	-0.135	0.224	1.000	5.000

Table B2 continued	Valid	Missing	Mean	Std. Deviation	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis	Minimum	Maximum
SRL-P-MA1	473	0	3.763	0.890	-0.878	0.112	0.798	0.224	1.000	5.000
SRL-P-MA2	473	0	3.332	1.082	-0.357	0.112	-0.804	0.224	1.000	5.000
SRL-P-MA3	473	0	3.744	0.866	-0.817	0.112	0.691	0.224	1.000	5.000
SRL-P-AD1	473	0	3.294	0.902	-0.248	0.112	-0.301	0.224	1.000	5.000
SRL-P-AD2	473	0	3.626	0.928	-0.642	0.112	0.137	0.224	1.000	5.000
SRL-P-AD3	473	0	3.218	1.090	-0.204	0.112	-0.971	0.224	1.000	5.000
SRL-P-AD4	473	0	3.135	1.014	-0.139	0.112	-0.760	0.224	1.000	5.000
SRL-P-AD5	473	0	3.641	0.935	-0.631	0.112	-0.156	0.224	1.000	5.000
SRL-P-AD6	473	0	3.323	0.961	-0.354	0.112	-0.475	0.224	1.000	5.000
SRL-P-MON1	473	0	3.841	0.825	-1.131	0.112	1.705	0.224	1.000	5.000
SRL-P-MON2	473	0	3.550	0.980	-0.615	0.112	-0.213	0.224	1.000	5.000
SRL-P-MON3	473	0	3.829	0.875	-0.976	0.112	1.001	0.224	1.000	5.000
SRL-P-ASE1	473	0	2.516	1.197	0.286	0.112	-1.094	0.224	1.000	5.000
SRL-P-ASE2	473	0	2.550	1.224	0.382	0.112	-0.919	0.224	1.000	5.000
SRL-P-ASE3	473	0	3.080	1.260	-0.254	0.112	-1.069	0.224	1.000	5.000
CH-MOT1	473	0	3.040	1.225	-0.007	0.112	-1.137	0.224	1.000	5.000
CH-MOT2	473	0	3.334	1.127	-0.354	0.112	-0.816	0.224	1.000	5.000
CH-MOT3	473	0	3.097	1.216	-0.038	0.112	-1.165	0.224	1.000	5.000
CH-GTM1	473	0	3.085	1.128	-0.060	0.112	-0.963	0.224	1.000	5.000
CH-GTM2	473	0	3.101	1.250	-0.121	0.112	-1.195	0.224	1.000	5.000
CH-GTM3	473	0	3.334	1.207	-0.252	0.112	-1.85	0.224	1.000	5.000
CH-GTM4	473	0	2.550	1.333	0.472	0.112	-1.038	0.224	1.000	5.000
CH-ISE1	473	0	3.894	1.122	-1.009	0.112	0.290	0.224	1.000	5.000
CH-ISE2	473	0	3.835	1.162	-0.880	0.112	-0.167	0.224	1.000	5.000
CH-ISE3	473	0	3.884	1.074	-0.890	0.112	0.020	0.224	1.000	5.000
CH-ISE4	473	0	3.649	1.103	-0.606	0.112	-0.508	0.224	1.000	5.000
CH-COG1	473	0	3.097	1.081	-0.134	0.112	-0.826	0.224	1.000	5.000
CH-COG2	473	0	3.146	1.078	-0.100	0.112	-0.876	0.224	1.000	5.000
CH-COG3	473	0	2.469	1.198	0.555	0.112	-0.654	0.224	1.000	5.000
CH-COG4	473	0	2.727	1.073	0.302	0.112	-0.793	0.224	1.000	5.000
CH-COG5	473	0	3.190	1.158	-0.179	0.112	-1.001	0.224	1.000	5.000
CH-COG6	473	0	2.928	1.108	0.124	0.112	-0.989	0.224	1.000	5.000
CH-S/EM1	473	0	3.710	1.091	-0.695	0.112	-0.282	0.224	1.000	5.000
CH-S/EM2	473	0	3.383	1.300	-0.371	0.112	-1.082	0.224	1.000	5.000
CH-S/EM3	473	0	3.723	1.117	-0.638	0.112	-0.475	0.224	1.000	5.000
CH-S/EM4	473	0	3.488	1.157	-0.429	0.112	-0.891	0.224	1.000	5.000
CH-S/EM5	473	0	3.002	1.238	0.084	0.112	-1.118	0.224	1.000	5.000
CH-S/EM6	473	0	3.655	1.203	-0.670	0.112	-0.576	0.224	1.000	5.000
CH-MET1	473	0	2.962	1.142	0.160	0.112	-1.014	0.224	1.000	5.000
CH-MET2	473	0	2.871	1.077	0.228	0.112	-0.980	0.224	1.000	5.000
CH-MET3	473	0	3.006	1.100	0.045	0.112	-1.092	0.224	1.000	5.000
CH-MET4	473	0	3.266	1.097	-0.214	0.112	-0.966	0.224	1.000	5.000
CH-MET5	473	0	2.922	1.120	0.227	0.112	-1.126	0.224	1.000	5.000
CH-MET6	473	0	2.956	1.064	0.099	0.112	-0.930	0.224	1.000	5.000
CH-MET7	473	0	2.922	1.192	0.189	0.112	-1.065	0.224	1.000	5.000
CH-MET8	473	0	3.118	1.061	-0.024	0.112	-0.954	0.224	1.000	5.000

Table B3
Factor-Level Pearson’s Correlations

Variable	1. Plan-IMP	2. MB-IMP	3. FAB-IMP	4. ASE--IMP	5. Plan-SE	6. MB_SE	7. Fab-SE	8. ASE-SE	9. TU-P	10. GM-P	11. TV-P	12. TM-P	13. MA-P	14. AD-P	15. MON-P	16. ASE-P	17. MOT-Ch	18. GTM-Ch	19. ISE-Ch	20. COG-Ch	21. SE-Ch	22. MET-Ch	23. GPA
1. Plan-IMP	—																						
2. MB-IMP	0.37***	—																					
3. FAB-IMP	0.383***	0.253***	—																				
4. ASE--IMP	0.163***	0.434***	0.282***	—																			
5. Plan-SE	0.506***	0.221***	0.365***	0.19***	—																		
6. MB_SE	0.327***	0.596***	0.301***	0.378***	0.439***	—																	
7. Fab-SE	0.274***	0.15**	0.731***	0.234***	0.475***	0.341***	—																
8. ASE-SE	0.142**	0.296***	0.232***	0.663***	0.308***	0.443***	0.324***	—															
9. TU-P	0.462***	0.339***	0.309***	0.161***	0.35***	0.364***	0.317***	0.243***	—														
10. GM-P	0.519***	0.376***	0.315***	0.223***	0.466***	0.417***	0.313***	0.286***	0.536***	—													
11. TV-P	0.117*	0.247***	0.048	0.208***	0.017	0.194***	0.019	0.167***	0.22***	0.149**	—												
12. TM-P	0.443***	0.193***	0.259***	0.09	0.445***	0.234***	0.309***	0.156***	0.371***	0.54***	0.126**	—											
13. MA-P	0.184***	0.328***	0.049	0.218***	0.108*	0.321***	0.111*	0.159***	0.288***	0.258***	0.387***	0.265***	—										
14. AD-P	0.194***	0.41***	0.102*	0.261***	0.101*	0.345***	0.119**	0.236***	0.227***	0.313***	0.221***	0.191***	0.37***	—									
15. MON-P	0.264***	0.352***	0.231***	0.217***	0.187***	0.337***	0.207***	0.192***	0.417***	0.386***	0.359***	0.244***	0.372***	0.235***	—								
16. ASE-P	0.152***	0.153***	0.217***	0.349***	0.168***	0.169***	0.24***	0.466***	0.234***	0.209***	0.143**	0.19***	0.155***	0.236***	0.142**	—							
17. MOT-Ch	-0.19***	-0.207***	-0.162***	-0.25***	-0.326***	-0.331***	-0.276***	-0.316***	-0.244***	-0.226***	-0.057	-0.18***	-0.047	-0.006	-0.201***	-0.112*	—						
18. GTM-Ch	-0.27***	-0.074	-0.247***	-0.093*	-0.516***	-0.202***	-0.404***	-0.184***	-0.22***	-0.305***	0.018	-0.383***	0	0.084	-0.102*	-0.032	0.466***	—					
19. ISE-Ch	-0.205***	-0.137**	-0.134**	-0.138**	-0.399***	-0.264***	-0.276***	-0.189***	-0.158***	-0.218***	0	-0.241***	-0.022	-0.05	-0.129**	-0.101*	0.519***	0.513***	—				
20. COG-Ch	-0.107*	-0.083	-0.152***	-0.108*	-0.276***	-0.238***	-0.192***	-0.202***	-0.336***	-0.204***	-0.084	-0.138**	-0.027	0.036	-0.184***	-0.092*	0.551***	0.466***	0.318***	—			
21. SE-Ch	0.001	-0.001	-0.08	-0.236***	-0.249***	-0.141**	-0.18***	-0.372***	-0.078	-0.104*	0.022	-0.182***	0.082	0	0.004	-0.153***	0.41***	0.368***	0.33***	0.348***	—		
22. MET-Ch	-0.292***	-0.123**	-0.214***	-0.087	-0.509***	-0.301***	-0.341***	-0.202***	-0.357***	-0.389***	-0.011	-0.372***	-0.071	0.007	-0.219***	-0.081	0.515***	0.589***	0.418***	0.618***	0.479***	—	
23. GPA	0.157***	-0.014	0.193***	-0.099*	0.178***	0.039	0.245***	0	0.231***	0.105*	0.017	0.117*	-0.019	-0.114*	0.082	0.061	-0.198***	-0.186***	-0.087	-0.249***	0.028	-0.226***	—

Note. * p < .05, ** p < .01, *** p < .001
Average scores are used to represent the factors

Table B4*Internal Consistency Reliability Estimates for the 4 scales*

	McDonald's ω	95% CI lower bound	95% CI upper bound	Mean	Sd
<i>SRL Importance (SRL-I)</i>	0.80	0.775	0.83	3.318	0.542
Planning	0.690	0.642	0.737	3.866	0.107
Metacognition & Beliefs	0.736	0.697	0.775	3.123	0.185
Foundational Academic Behaviours	0.635	0.578	0.693	3.803	0.717
Academic Social Engagement	0.769	0.734	0.805	2.541	0.249
<i>SRL Self-efficacy (SRL-SE)</i>	0.841	0.820	0.862	3.023	0.609
Planning	0.768	0.733	0.804	3.216	0.197
Metacognition & Beliefs	0.777	0.745	0.810	2.940	0.089
Foundational Academic Behaviours	0.696	0.648	0.743	3.538	0.895
Academic Social Engagement	0.783	0.749	0.817	2.426	0.149
<i>SRL Practices (SRL-P)</i>	0.873	0.857	0.890	3.523	0.456
Goal Management	0.843	0.821	0.865	3.513	0.247
Task Understanding	0.731	0.692	0.769	3.874	0.148
Task Value	0.718	0.675	0.762	3.540	0.182
Motivation Appraisal	0.667	0.625	0.710	3.613	0.244
Adapting	0.784	0.754	0.814	3.371	0.212
Monitoring	0.790	0.757	0.822	3.738	0.165
Academic Social Engagement	0.800	0.769	0.831	2.711	0.316
Time Management	0.646	0.608	0.685	3.730	0.274
<i>SRL Challenges (SRL-C)</i>	0.930	0.921	0.939	3.204	0.654
Metacognitive	0.878	0.862	0.895	3.003	0.128
Social & Emotional	0.834	0.810	0.857	3.488	0.274
Cognitive	0.834	0.811	0.857	2.928	0.279
Initiating & sustaining engagement	0.780	0.748	0.812	3.815	0.115
Goal & Time management	0.810	0.782	0.838	3.007	0.319
Motivation	0.682	0.620	0.745	3.157	0.156

Table B5*Item-Factor Loadings for the SRL-Importance and SRL-Self-Efficacy Scales*

	SRL-IMPORTANCE (SRL-I)			SRL-SELF-EFFICACY (SRL-SE)		
	β	St Error	p	β	St Error	p
Planning						
Thoroughly understanding what I need to do	0.557	0.036	< .001	0.557	0.044	< .001
Managing my goals	0.742	0.042	< .001	0.805	0.041	< .001
Managing my time	0.638	0.045	< .001	0.792	0.045	< .001
Metacognition & Beliefs						
Changing my beliefs or approaches as needed	0.517	0.046	< .001	0.659	0.042	< .001
Thinking about my effort and feelings related to this study session	0.647	0.048	< .001	0.649	0.043	< .001
Finding value or usefulness in the study session	0.63	0.046	< .001	0.736	0.04	< .001
Monitoring my learning	0.734	0.042	< .001	0.678	0.042	< .001
Foundational Academic Behaviours						
Showing up for class or online meetings	0.613	0.057	< .001	0.615	0.056	< .001
Getting myself to do schoolwork on time	0.521	0.053	< .001	0.785	0.051	< .001
Participating in learning activities or discussions	0.653	0.056	< .001	0.556	0.055	< .001
Academic Social Engagement						
Having fun and enjoying academic work	0.608	0.054	< .001	0.7	0.047	< .001
Making friends and getting to know people	0.739	0.053	< .001	0.717	0.052	< .001
Feeling like I belong to my program and university	0.815	0.047	< .001	0.779	0.047	< .001

Table B6*Correlation between Factors of SRL-Importance and SRL-Self Efficacy Scales*

SRL-Importance					SRL-Self Efficacy			
	Plan-IMP	MB-IMP	FAB-IMP	ASE-IMP		Plan-SE	MB-SE	FAB-SE ASE-SE
Plan-IMP	1				Plan-SE	1		
MB-IMP	0.535***	1			MB-SE	0.567***	1	
FAB-IMP	0.557***	0.415***	1		FAB-SE	0.677***	0.433***	1
ASE-IMP	0.22**	0.548***	0.404***	1	ASE-SE	0.38***	0.575***	0.406***

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

Table B7
Item-Factor Loadings for the SRL-Practices Scale

	β	St Error	<i>p</i>
Task Understanding			
Asked myself if I know what is important to learn	0.513	0.044	< .001
Made sure I understand terminology used in instructions	0.58	0.034	< .001
Thought about the professor's expectations and standards	0.532	0.039	< .001
Considered what knowledge or big ideas I should learn or demonstrate	0.729	0.036	< .001
Thought about relevant information and resources	0.617	0.040	< .001
Goal Management			
Set goals for my work	0.697	0.039	< .001
Made goals to learn, understand, or remember	0.794	0.040	< .001
Set goals that will be useful for checking my own progress	0.748	0.045	< .001
Identified specific content, ideas, or terms in my goals	0.678	0.041	< .001
Assessed my goal attainment	0.672	0.044	< .001
Task Value			
Thought about why we are being asked to know this stuff	0.64	0.054	< .001
Reflected on why this work is important	0.767	0.046	< .001
Made judgments about the usefulness or value of the content	0.632	0.042	< .001
Time Management			
Checked to see if I am staying on time	0.607	0.046	< .001
Chose goals that can be completed in one reasonable work session	0.531	0.056	< .001
Created a timeline or schedule	0.651	0.058	< .001
Motivation Appraisal			
Assessed if I think I can do it	0.623	0.047	< .001
Assessed my feelings for the task	0.614	0.057	< .001
Evaluated the effort I was putting in	0.57	0.046	< .001
Adaptation			
Changed my understanding of the task at hand	0.583	0.044	< .001
Modified my plans for the task	0.658	0.052	< .001
Switched to a different strategy or approach	0.664	0.047	< .001
Changed my feelings about the task	0.611	0.045	< .001
Altered the level of effort I put in	0.541	0.045	< .001
Modified my beliefs about how well I would do my tasks	0.605	0.043	< .001
Monitoring			
Asked myself if I understand what I am supposed to be doing	0.682	0.037	< .001
Asked myself if I was remembering	0.737	0.043	< .001
Asked myself if I am understanding the material	0.809	0.037	< .001
Academic Social Engagement			
Got to know people in the class	0.83	0.052	< .001
Had fun in online chats, discussions or via email with the instructor or other students	0.768	0.054	< .001
Helped fellow students	0.671	0.056	< .001

Table B8
Corelation between Factors of SRL-Practices Scale

	TU-P	GM-P	ATV-P	MA-P	TM-P	AD-P	MON-P	ASE-P
TU-P	1							
GM-P	0.672***	1						
ATV-P	0.282**	0.178*	1					
MA-P	0.533***	0.457***	0.615***	1				
TM-P	0.539***	0.732***	0.197*	0.482***	1			
AD-P	0.312***	0.387***	0.305***	0.618***	0.286**	1		
MON-P	0.526***	0.461***	0.466***	0.635***	0.332**	0.297***	1	
ASE-P	0.305***	0.243***	0.186**	0.291***	0.254**	0.311***	0.187**	1

*Note. Residual covariance between MA1 and MA2 : 0.309***.*

*Residual covariance between AD5 and AD6 : 0.28***.*

p < .05. **p < .01. *p < .001*

Table B9*Factor Loadings for the Academic Challenges Scale*

	β	Std Error	<i>p</i>
<i>Motivation Challenges</i>			
Believing I can do my work	0.633	0.054	< .001
Feeling like my work was worth doing	0.637	0.052	< .001
Persisting when things got tough	0.629	0.054	< .001
<i>Cognitive Challenges</i>			
Applying my learning	0.713	0.045	< .001
Elaborating and connecting what I learned	0.751	0.045	< .001
Grappling with language and communication	0.519	0.054	< .001
Identifying important ideas	0.702	0.045	< .001
Remembering things I learned	0.635	0.050	< .001
Understanding learning materials and concepts	0.739	0.046	< .001
<i>Metacognitive Challenges</i>			
Breaking tasks down	0.687	0.048	< .001
Interpreting tasks and expectations	0.68	0.045	< .001
Knowing how to tackle tasks	0.739	0.045	< .001
Knowing how or when to fix strategies/study skills	0.664	0.047	< .001
Prioritizing or figuring out what is important	0.658	0.048	< .001
Monitoring my task or learning progress	0.731	0.044	< .001
Setting goals for studying	0.644	0.045	< .001
Applying the right strategies/techniques	0.696	0.051	< .001
<i>Social & Emotional Challenges</i>			
Feeling connected	0.671	0.049	< .001
Feeling lonely	0.652	0.058	< .001
Finding enjoyment at university	0.635	0.048	< .001
Managing my emotions/feelings	0.747	0.056	< .001
Managing relationships	0.575	0.050	< .001
Taking care of my mental health and well-being	0.767	0.049	< .001
<i>Initiating & Sustaining Engagement Challenges</i>			
Wanting to do my schoolwork	0.715	0.050	< .001
Putting things off (procrastinating)	0.61	0.053	< .001
Focusing or concentrating	0.784	0.046	< .001
Managing distractions	0.642	0.049	< .001
<i>Goal & Time Management Challenges</i>			
Keeping my commitments or goals	0.695	0.048	< .001
Managing my time, tasks, or goals	0.811	0.052	< .001
Meeting deadlines	0.582	0.060	< .001
Organizing time and tasks	0.786	0.055	< .001

Table B10
Corelation between Factors of Academic Challenges Scale

	MOT-CH	COG-CH	MET-CH	S/EM-CH	ISE-CH	GTM-CH
MOT-CH	1					
COG-CH	0.772***	1				
MET-CH	0.766***	0.725***	1			
S/EM-CH	0.676***	0.411***	0.557***	1		
ISE-CH	0.649***	0.406***	0.5***	0.427***	1	
GTM-CH	0.726***	0.547***	0.678***	0.447***	0.669***	1

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

Appendix C
Supplementary Statistics related to the Structural Models

Table C1

Correlation among Predictors of the Concurrent-Predictor Model Presented in Table 5

	Plan-I	MB-I	FAB-I	ASE-I
Plan-I	1			
MB-I	0.536***	1		
FAB-I	0.571***	0.413***	1	
ASE-I	0.219***	0.548***	0.387***	1

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

Table C2

Correlation among Predictors of the Concurrent-Predictor Model Presented in Table 6

	Plan-SE	MB-SE	FAB-SE	ASE-SE
Plan-SE	1			
MB-SE	0.567***	1		
FAB-SE	0.677***	0.429***	1	
ASE-SE	0.38***	0.574***	0.403***	1

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

Table C3

Correlation among Predictors of the Concurrent-Predictor Model Presented in Table 7

	TU-P	GM-P	ATV-P	MA-P	TM-P	AD-P	MON-P	ASE-P
TU-P	1							
GM-P	0.672***	1						
ATV-P	0.282***	0.178**	1					
MA-P	0.533***	0.457***	0.615***	1				
TM-P	0.539***	0.732***	0.197**	0.482***	1			
AD-P	0.312***	0.387***	0.305***	0.618***	0.286***	1		
MON-P	0.526***	0.461***	0.466***	0.635***	0.332***	0.297***	1	
ASE-P	0.305***	0.243***	0.186**	0.291***	0.254***	0.311***	0.187**	1

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

Table C4

Correlation among Predictors of the Concurrent-Predictor Model Presented in Table 8

	MOT-CH	COG-CH	MET-CH	S/EM_CH	ISE-CH	GTM-CH
MOT-CH	1					
COG-CH	0.771***	1				
MET-CH	0.765***	0.725***	1			
S/EM_CH	0.675***	0.412***	0.558***	1		
ISE-CH	0.648***	0.405***	0.5***	0.427***	1	
GTM-CH	0.727***	0.547***	0.678***	0.447***	0.668***	1

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

Table C5*Correlation among Predictors of the Model Presented in Figure 1*

	TU-P	GM-P	ATV-P	MA-P	TM-P	AD-P	MON-P	ASE-P
TU-P	1							
GM-P	0.676***	1						
ATV-P	0.292***	0.186**	1					
MA-P	0.533***	0.474***	0.604***	1				
TM-P	0.534***	0.755***	0.155*	0.463***	1			
AD-P	0.298***	0.389***	0.307***	0.612***	0.297***	1		
MON-P	0.534***	0.461***	0.467***	0.638***	0.329***	0.304***	1	
ASE-P	0.301***	0.241***	0.196***	0.289***	0.223**	0.323***	0.186**	1

Standardized Residual Covariance among Outcome Variables of the Model Presented in Figure 1

	MOT-CH	COG-CH	MET-CH	S/EM_CH	ISE-CH	GTM-CH
MOT-CH	1					
COG-CH	0.772***	1				
MET-CH	0.721***	0.721***	1			
S/EM_CH	0.684***	0.439***	0.577***	1		
ISE-CH	0.621***	0.399***	0.426***	0.38***	1	
GTM-CH	0.71***	0.599***	0.571***	0.413***	0.63***	1

Note. * $p < .05$. ** $p < .01$. *** $p < .001$

Appendix B: Study 2

Study 2: Academic Challenges under the Lens of Item Response Theory: An Item Level Analysis of Measurement Precision in Perceived Academic Challenges Scale

Ramin Rostampour

Department of Educational Psychology and Leadership Studies

University of Victoria

This research was supported by a Social Sciences and Humanities Research Council (SSHRC) of Canada multisite partnership grant (the UniVenture consortium; PI Sherry Stewart, 895-2019-1021).

Abstract

This study employed Item Response Theory (IRT) to validate a measure of academic challenges, examining a sample of 2671 university students enrolled in five North American universities. The instrument is designed to measure six dimensions of perceived academic challenges related to the regulation of: *motivation, initiating and sustaining engagement, goal and time management, cognitive, metacognitive, and social and emotional* aspects. These dimensions correspond to students' struggles in key self-regulatory areas, which are closely linked to success in self-regulated learning. The current IRT analyses expanded on earlier factor analysis validation efforts by examining the precision with which items constituting the measure assess the intended constructs. Considering that one of the main objectives of the measure is its practical application in identifying students who judge they are struggling, the study also evaluated the psychometric properties of composite scores, which are derived through straightforward arithmetic computations. The study also investigated how the measure is related to students' GPA expectations, self-efficacy for GPA expectations, and actual GPA at the end of the semester. Findings confirmed the measure's adequate psychometric properties, the adequacy of its total score, as well as its relevance to the criterion constructs. Additionally, recommendations for improving the measure were presented. This study underscores the critical need to refine self-reported measures of students' intentions, perceptions, and experiences as an important modality in the broader multimodal understanding of learning.

Background

In their educational journey, students frequently experience a range of difficulties that influence their learning process. These perceived challenges, spanning motivational, behavioural,

cognitive, metacognitive, social, and emotional dimensions, are critical to identify and comprehend from the students' viewpoint.

The current study aims to enhance previous research by continuing the validation process of a measure that assesses perceived academic challenges among university students. This scale is anchored in the Self-Regulated Learning (SRL) framework.

Despite theoretical differences among several SRL models, they all define SRL as a relatively independent type of learning guided by metacognition, and characterized by active participation, and strategic action (e.g., planning, monitoring, enacting, and evaluating), and involving intentional regulation of diverse range of overt and covert behaviours (Dunlosky & Ariel, 2011; Hadwin et al., 2017; Hadwin & Winne, 2012; Kim et al., 2020; Schunk & Greene, 2018; Winne, 1997; Zimmerman, 2013). In other words, SRL involves judgment and modulation across various domains in which individuals engage to leverage and adapt their personal and environmental characteristics to achieve desired learning outcomes. The regulatory domains explored and enhanced within SRL encompass cognitive, metacognitive, behavioral, motivational, and affective domains (see Panadero, 2017, for a review of six SRL models).

A cornerstone of SRL as a proactive mode of learning is the development of adaptive beliefs, such as the growth mindset (Dweck, 2015, 2016; Yeager & Dweck, 2020) which frames learning and self-regulation as an ongoing, improvable process rather than a fixed ability. According to Zimmerman (2015), self-regulated learning is neither an aptitude nor a mere performance skill. Instead, it is a transformative process that turns personal capacities into performance skills. This perspective encourages learners to focus not only on the end results of their learning but also on understanding the nuances of the learning process itself, fostering a deeper comprehension of how they learn and paving the way for ongoing academic growth

(Cambridge, 2013; Cleary, 2004; Pino-Pasternak, et al., 2014; Whittlesey 2021; Zimmerman, et al., 2011).

In addition to an understanding of the learning processes and task requirements, such an improvement requires learners to develop an awareness of their personal preferences, capabilities, strengths, and challenges. By cultivating such a self-awareness, learners can tailor their strategies to fit their specific needs, thus enhancing the effectiveness of their study behaviours.

Winne defines such a personal theory of learning in terms of an IF-THEN-ELSE model by which learners (a) form hypotheses about their own learning, (b) actively test these hypotheses through experimentation, (c) reflect on the outcomes, and (d) make adaptive revisions. This dynamic process enables the transfer of learning to new, albeit related, tasks (Winne, 1997, 2006, 2010a, 2017, 2022). Challenges can be viewed as key representations in such a personal theory of learning that are shaped in relation to retrospective experiences and current and prospective goals.

Academic Challenges

While there is a wealth of research on the self-regulatory behaviours that contribute to student success, there is a paucity of studies exploring the specific challenges students face while studying. The academic challenges scale was designed to address this gap (Hadwin et al., 2024). Existing findings suggest time management, cognitive challenges, concentration, and boredom are the most common challenges reported by students (Koivuniemi, et al., 2017). A study by Hadwin et al. (2019) found that motivation and emotional challenges, goal setting and planning, and cognitive challenges were the most frequently reported weekly challenges over an academic semester.

The perception of a challenge is created when students become aware of a discrepancy between a desired state and the current state. Being challenged by a task can be a functional trigger of adaptive behaviour that prompts students to re-evaluate their current inadequate approach and engage in a plausibly more effective phase of self-regulation to close this discrepancy (Hadwin et al., 2019). To elaborate, from a self-regulatory perspective, encountering challenges is not just a setback, but a crucial situation for developing more sophisticated and effective learning strategies. However, if a student's regulatory efforts are ineffective, especially in the face of intense and prolonged challenges, the functional role of challenges can shift from being opportunities for growth to barriers that impede learning and engagement (Boekaerts, 2011; Tuominen-Soini, et al., 2012; Zimmerman & Cleary, 2009). Such a functional shift underscores the need for timely identification of challenges to help students transform challenges into opportunities for learning and development.

SRL as goal driven activity is often viewed as a function of two broad metacognitive skills (a) monitoring: identifying discrepancy between current and desired states and (b) control: overt or covert actions to reduce the discrepancy (Dunlosky & Ariel, 2011; Schunk & Greene, 2018; Winne, 1997, 2001, 2005 2017a; Winne & Hadwin 1998; Zimmerman, 2013). Metacognitive knowledge and metacognitive awareness, although not sufficient factors for successful metacognitive control, seem to be essential foundations upon which intentional control processes rest (for a discussion on automaticity of SRL regulatory processes, See Perry & Winne, 2006).

A measure of academic challenges can be an informative self-assessment tool for students. When coupled with personalized feedback, this self-assessment transforms into a self-reflection mechanism, setting a stage for enhancing students' metacognitive awareness of their

perceived challenges. This increased awareness becomes even more beneficial when it is complemented with psychoeducational resources. These resources assist students in comprehending their self-assessment feedback, encourage the adoption of suitable learning strategies based on this feedback, and guide them towards seeking customized support, if necessary. Consequently, students are better prepared to strategically address their challenges, resulting in a more an improved overall learning experience.

Furthermore, measuring academic challenges can offer valuable insights for institutional planning and classroom diagnostics. It provides a structured overview of the difficulties students face, enabling educators and institutions to extend their focus beyond mere academic performance. This deeper insight into students' struggles is pivotal for devising focused interventions that cater to specific needs rather than relying on general assumptions. For instance, identifying students who are demotivated can lead to targeted initiatives, such as workshops or counseling, aimed at enhancing motivation and engagement, boosting retention rates, among other things. Similarly, applying this measure in particular courses can reveal predominant challenges, whether cognitive, motivational, or otherwise, facilitating course adjustments to better serve student needs.

In research contexts, measuring perceived academic challenges offers valuable insights into learning behaviors and beliefs by situating them within the specific challenges encountered. This approach allows for an understanding of the behaviors as responses to these challenges, thereby providing a comprehensive view of how students manage their learning. This dual-purpose design allowing for research and self-assessment use is based on the underlying assumption that the nature of the data required to understand and enhance management of

learning is the same for both researchers and learners (Butler, 2018; Hadwin et al., 2019, Winne, 2017b).

Academic Challenges Scale

The Academic Challenges scale is one of the four measures in the Self-Regulated Learning Profiles and Self-Diagnostic tool (SRL-PSD, Hadwin et al., 2024). The other three psychometrically separate scales measure students' perceptions of the importance of self-regulatory practices (SRL-importance), their confidence in employing these practices (SRL-self-efficacy), and their self-reported rates of deployment of self-regulatory practices (SRL-Practices). The concurrent use of the four measures provides a comprehensive and relational profile of students' self-regulated learning (SRL-PSD, Hadwin et al., 2024).

The Academic Challenges scale, presented in the supplementary materials A, was iteratively developed over several years of data collection and pilot testing with post-secondary students. Initially, Hadwin et al. (2019) asked students to document the challenges encountered during their weekly study sessions throughout a semester, aiming to capture direct insights from the students' experiences. Subsequently, the responses were analyzed qualitatively and organized into a dropdown list, offering a structured method for subsequent students to identify encountered challenges. An *other* option was included, providing an open text field for contributions of any unlisted challenges to ensure the list's comprehensiveness in reflecting students' actual experiences. This list was refined and expanded until a saturation point was reached, characterized by a negligible rate of uncategorized challenge types being reported over two consecutive semesters.

Subsequently, it was hypothesized that the challenges could be categorized into five principal targets for regulation linked to student success: motivation, behavior, cognition,

metacognition, and social-emotional challenges. This classification was designed to comprehensively represent the varied academic challenges students reported. The first version of the academic challenges scale was developed based on this conceptual model, and its psychometric properties were tested using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) with a two samples of university students (Hadwin et al., 2024).

An EFA analysis was conducted after the *a priori* 5 factor solution was not supported by a confirmatory factor analysis. EFA revealed that two items from the behavioral dimension, along with one each from the motivation and cognitive dimensions, coalesced to represent challenges with procrastination and managing distractions. This scale was named *Initiating and Sustaining Engagement (ISE)*. The remaining items of the former behavioral challenges scale represented challenges in goal setting and time management. Therefore, the behavioural challenges factor was renamed *Goal and Time Management (GTM)* to better reflect the content of items. The reconfigured measurement model for the scale included 6 subscale and showed acceptable psychometric properties. The convergent validity of the scale in relation to the other measures of SRL-PSD, and its criterion validity in predicting end of semester GPA was also established (Hadwin et al., 2024).

Rational for the Current Study

As mentioned earlier, the initial investigation of the academic challenges scale employed factor analysis (EFA & CFA) and supported adequate psychometric quality of the scale and its criterion validity for predicting academic performance, as measured by GPA (Hadwin et al., 2024).

The objective of the current study was to use Item Response Theory (IRT) as a complementary method to further investigate the quality of the items that constitute the academic challenges scale.

CFA is a powerful tool for investigating the validity of inferences about the match of items to constructs, primarily by examining the relationships between observed items and the latent constructs underlying the observations (Kline, 2015). It affords insights based on the pattern of item loadings on their corresponding latent factors, the correlations among latent factors, and the proportion of variance in the items that can be attributed to latent factors. A key distinction between IRT and CFA lies in their focus. While CFA provides a global perspective, assessing how well a hypothesized model aligns with empirical data, IRT offers a more specific item-level analysis.¹ In the remainder of this section, the added value of employing IRT in the current study compared to the earlier validation study is described.

Evaluating Measurement Quality Across Latent Attribute Levels. IRT extends measurement modeling beyond test characteristics to afford assessments of psychometric validity in terms of person-characteristics (Embretson & Reise, 2013). To elaborate, the assumptions about measurement quality differ between the two methods. CFA operates under the assumption that the quality of measurement is uniform for all individuals. IRT, on the other hand, posits that measurement quality can vary depending on individuals' level of the latent attribute² under

¹ Similar objectives could be achieved by examining local parameter estimates in a CFA, such as item-factor loadings, residual errors, and model-based internal consistency indices. However, evaluating the global fit of a theoretical model to data is the primary objective of a confirmatory factor analysis.

² In the context of this study, the term "latent attribute" is used to denote the underlying construct that the academic challenges subscales aim to measure. This terminology was selected to move away from more commonly used terms like "ability" or "trait," which do not accurately represent the essence of academic challenges.

investigation (Embretson & Reise, 2013). To clarify, IRT models the probability of endorsing a response is a function of both person parameters (one's level of the latent attribute) and item parameters (e.g., item thresholds and discrimination). To exemplify, certain items within the motivation challenges scale might offer more precise measurements (i.e., have a lower standard error of measurement) for individuals with an average level of motivation challenges than for those with very high or low levels. The values of item-level parameters can be aggregated and reported for the entire scale, indicating the extent to which a person's level of motivational challenges can be inferred from their response to certain items, and the extent to which the scale comprises items that provide information across a wide range of motivational challenges. Given the diagnostic objectives of the scale, understanding the distribution range of academic challenges that the scale can measure reliably is essential. If necessary, additional items can be included to cover areas that are under-measured by the scale.

Evaluating the Quality of the Response Scale. Beyond helping to understand how effectively the items can distinguish between individuals with differing levels of that attribute (e.g., low, average, high), and what levels of the latent attribute are needed to endorse a response, IRT enables assessing the quality of the ordinal response options categories in differentiating between individuals (Ostini & Nering, 2006). This feature is particularly useful in evaluating the quality of the response scale used for the academic challenges scales, which employs a 5-point Likert scale. For instance, if there is excessive overlap between two categories, it may suggest the potential to merge them into a single response category. On the other hand, large gaps between response categories could signal a need for modification, ensuring they capture a wide enough range of responses at various levels of the latent attribute.

Visual Representation of Measurement Properties One of the significant benefits of employing IRT, particularly with unidimensional IRT models, lies in its capacity to graphically depict the measurement quality. This visual representation facilitates a more straightforward understanding and communication of the measurement's quality, making it easier to identify underperforming items. This process lays the foundation for enhancing the quality of scales through careful item revision or, when required, the removal of problematic items.

Developing a Shorter Version of the Scales. Similarly, IRT can offer valuable insights into the feasibility of developing a shorter version of the scales by identifying items that yield the most information across broad spectrum of the latent attribute. Such a refinement can suggest ways to streamline the scales without significantly compromising their psychometric validity and reliability. Despite being brief, these scales effectively gather essential information, making them ideal for frequent data collection studies, like experience sampling methods, where minimizing participant burden is key.

Evaluating Differential Item Functioning. Differential Item Functioning (DIF) in IRT facilitates identifying items where individuals belonging to different demographic groups (e.g., first year/second year students; domestic/international, etc.) demonstrate differing likelihoods of providing specific responses to an item even if they possess the same level of underlying attribute that the scale is designed to measure. (Osterlind & Everson, 2009).

Evaluating the Psychometric Properties of Each Subscale as a discrete Scale. Structural Equation Modeling (SEM) has become a prevalent analytic framework in social sciences and psychology (Karimi & Meyer, 2014; Tarka, 2018). However, due to avoiding overparameterization and convergence issues, and preserving statistical power, it may not always be feasible to include all subscales of a scale in the model, especially when the scale includes

multiple factors and the research questions do not directly pertain to all those specific factors. Consequently, a single factor or a subset of factors are often included and estimated in the analyses. Given that most multidimensional constructs are validated using a correlated factor CFA, information about the global fit is often taken as proof of the construct validity of a scale. The concept of fit represents a harmony between covariance structure of the data and the one reproduced by the specified model. However, the overall psychometric validity of a multidimensional latent model does not necessarily support the local validity of its subscales. Unidimensional IRT allows for examining the quality of measurement when each subscale is modeled as an independent construct.

Evaluating the Psychometric Adequacy of the Overall Scale as a Unidimensional Scale.

A prior validation study revealed that the six subscales of academic challenges exhibited moderate to high correlations with one another (Hadwin et al., 2024). Such a correlation among the subscales suggests that, although they aim to capture distinct dimensions of a broader construct, these dimensions are not entirely independent and instead share common underlying themes. High correlations might raise concerns regarding discriminant validity, indicating the subscales may not be sufficiently distinct. However, within the academic challenges context, such a covariance structure could imply that the subscales may collectively represent an overarching construct of general academic challenges. Considering the interconnected nature of academic challenges, it becomes pertinent to question the feasibility of deriving a unidimensional score to represent a general factor of academic challenges. Conducting an item-level analysis on this general unidimensional scale, encompassing all items, aids in evaluating the scale's measurement quality and internal structure regarding its constituent items. This analysis allows for proposing methods to compute the overall score for the scale in practical

settings. Additionally, it paves the way for the identification of measurement models that more accurately reflect the construct's multidimensionality. In an ongoing study, authors are exploring the concurrent modeling of specific and general factors of academic challenges using measurement models such as higher-order and bifactor exploratory structural equation modeling (ESEM, see Asparouhov & Muthen, 2009). Nevertheless, the feasibility of deriving an overall score for the scale for practical use warrants further examination through classical test theory and an item-level analysis, such as IRT.

Evaluating the Psychometric Adequacy of Composite Scores for the Scale. The academic challenges scale has been developed to fulfill both research and practical needs in educational settings. In practical settings like classrooms where the scale is used as a screening tool, latent measurement modeling may not be readily available. Factor indeterminacy also presents a challenge for computing factor scores for covariance-based structural equation models (Rigdon, 2019). Consequently, non-refined composite scores such as sum or average scores (DiStefano et al., 2019) are frequently used as substitutes for latent factors. This approach, however, often applies restrictive assumptions on modeling measurement error, such as assigning equal weights to contributing items in calculating average scores, or disproportionately allocating greater weights to scales with more items when computing overall sum scores. Also, researchers may be interested to use a regression-based approach to conduct conditional process analysis (Hayes & Rockwood, 2020, 2021) for which computing composite scores for the scales is necessary. The psychometric validity of a model derived from an analytic procedure such as CFA that explicitly accounts for measurement error cannot be directly used as an indication of validity of composite scores (McNeish & Wolf, 2020). Certain conditions, such as unidimensionality and high internal consistency of factors, must be met for composite scores to yield relatively consistent results

with covariance-based estimates of factor analysis and structural equations modeling (SEM) (Widaman & Revelle, 2023). Further analyses such as what was done in this study are needed to comprehend the extent to which composite scores deviate from or aligned with true factors.

This study was described as an effort to triangulate the psychometric quality of the academic challenges scale using a different analytical method. This cross-validation approach offers several benefits peripheral to the major ones mentioned earlier.

The differing ways in which factor analysis and IRT handle missing data serve as an illustration of the value of such a cross-validation. In IRT, missing responses are simply not included in the likelihood function, effectively bypassing the issue. On the other hand, CFA requires the implementation of methods such as Full Information Maximum Likelihood (FIML) or multiple imputation to address missing data (Finch, 2008). Alignment of these two methods for handling missingness tests the robustness of findings despite the presence of missing information. This difference in handling missing data is just one example of how the use of different analytic methods can provide a deeper understanding of the measurement quality of the scales.

IRT can be used as a prerequisite for developing *computerized adaptive scales* (Boyd et al., 2011). In such scales, rather than presenting each participant with the same set of items, the items are dynamically targeted based on the respondent's prior responses. This adjustment is guided by a theoretically informed and empirically fine-tuned underlying algorithm. As discussed, the quality of measurement in IRT is modeled as a function of both person and item characteristics. The estimation of each parameter is independent of the other, a phenomenon referred to as parameter invariance or sample invariance in IRT (Dodd et al., 1995). Such an invariance means that, given certain conditions are met, the measurement parameters are

unbiased and independent from sample characteristics and can be reasonably applied across new samples from the same population. These conditions include having a representative sample, random missingness, ensuring adequate coverage of the range of the latent attribute by the items, and having sufficient information about differential item functioning across demographic characteristics (Boyd et al., 2011).

The prospect of creating a computerized adaptive scale for measuring academic challenges, particularly with the aim of generating self-diagnostic personalized profiles, holds considerable promise. One area where various SRL measurements diverge is the level of granularity at which SRL should be measured (Winne, 2010b). Adaptive scales can provide opportunities to make an individualized balance between the scope (range) and depth (granularity) of measurement. This can be achieved by basing the measurement procedure on respondents' answers to a set of screening items and adaptively continuing the process until the model generates a desired broad and detailed profile of the individual's academic challenges. Just to give a simplified example, if a student reports relatively high levels of motivational challenges, the algorithm could present items that explore deeper into understanding the specifics of these challenges. This approach would be more effective than having the student respond to detailed items about an area where they declared they do not perceive much challenge.

To conclude, item-level analyses, such as Item Response Theory, can supplement CFA by facilitating a detailed analysis at the item level and by estimating measurement precision across the spectrum of the latent attribute. IRT can help evaluate the quality of the response options, identify underperforming items and items with differential functioning across groups, aid in creating a more concise version of the scale, and serve as an initial step towards developing computerized adaptive scales for measuring academic challenges. Additionally, this

study will explore the relationships between individual subscales of the academic challenges scales, as well as their overall composite scores, with three key criterion constructs: students' end-of-semester GPA, their self-reported expected GPA early in the semester, and their self-efficacy for achieving their expected GPA.

Objectives and Main Research Questions

The objective of the current study was to use item response theory as a complementary method to further investigate the quality of the items that constitute the academic challenges scale. The main research questions of the study are:

Research Question 1. How precisely do the items within each subscale of the academic challenges scale measure different levels of their target challenge as latent unidimensional constructs?

Research Question 2. How precisely do all items combined in a single scale measure different levels of a latent general academic challenges construct?

Research Question 3. How do estimated average and factor scores for individual subscales, and for the overall scale align with their corresponding true factors?

Research Question 4. How are individual subscales, the overall scale, and composite scores of the academic challenges scale associated with students' GPA expectations, GPA Self-Efficacy, and end-of-semester GPA?

Method

Participants and Data Collection Procedure

The data for the study were collected as part of the UniVenture program (Stewart et al., 2021), a wellness program tailored for university students. The program focused on offering specialized workshops that equipped students with valuable skills, facilitate the development of

long-term goals, and guide them in leveraging their unique personalities to effectively pursue their goals.

The current study used baseline data collected during weeks one to five of the Fall 2021 and Winter 2022 semesters via REDCap electronic data capture tools (Harris et al., 2019). The data comprise university students from five Canadian universities across four provinces (Nova Scotia, Ontario, Quebec, and British Columbia). Institutional data, such as GPA, was obtained from the universities' registrar's offices. The sample consisted of $N=2671$ participants, with 65% being female, 32% male, and 3% non-binary. Among the participants, 56% were first-year students, while 44% were in their second year of study. Students' age ranged from 18 to 25 (Mean=19, SD=1.5).

During the baseline phase of data collection, which served as the data source for the current study, students provided responses to a range of topics including demographics, personality, substance and alcohol use, anxiety, and depression prior to the academic challenges items.

All analyses were subjected to measurement invariance testing to guarantee uniformity across the participating universities, as well as to ensure consistency between the two cohorts under study: Fall 2021 and Winter 2022 intakes. After the data collection at baseline, about ten percent of students attended the program workshops. Given the potential influence of workshop attendance on the GPAs collected at the end of academic semesters, a sensitivity analysis was performed to confirm there was no statistically detectable difference in GPA between students who attended the workshops and those who did not.

Ethics

The study is funded through Social Sciences and Humanities Research Council (SSHRC) Partnership Grant [Grant#: 895-2019-1021]. The study is ethically approved by the five participating data collection sites by their respective ethic review boards [Dalhousie University (2021-5689), Saint Francis Xavier University (ROMEO File #25486), University of Montreal (2020-883), York University (2021-272) and University of British Columbia – Okanagan (H21-02007)]. Students' consent to access their institutional data from the registrar's office of their universities was obtained separately from their general consent to participate in the study. All non-consenting students were removed from the dataset. The dataset was fully anonymized, cleaned, and shared across sites after evaluating all ethical considerations and the existing regulations enforced by the relevant Research Ethics Boards.

Measures

Academic Challenges Scale

The Academic Challenges scale consists of 31 items spread across six subscales, each assessing the extent to which students faced various challenges in their studying over the past two weeks. The academic challenge scale asks students to indicate their level of agreement with experiencing challenges in these six areas over the past two weeks of study: (1) *motivation*, (2) *initiating and sustaining engagement*, (3) *goal and time management*, (4) *cognitive*, (5) *metacognitive*, and (6) *social and emotional challenges*. The response scale employed was a 5-point Likert scale, ranging from *strongly disagree* (1) to *strongly agree* (5), with options for *disagree* (2), *neutral* (3), and *agree* (4) in between.

The two-week timeframe was chosen to achieve a balance between accurately capturing recent study experiences and allowing a broad enough range of recallable incidents. Prompting

students to reflect on how various challenges characterize their recent study behavior, rather than their overall persona as a learner implies these malleable behaviours are typical of their studying over a particular timeframe and within a particular rather than a semester- or life-long circumstance rather than typical of their fixed traits as individual learners. Anchoring responses with respects to such a time-window constraint is based on two main considerations. First, while learners' perceived challenges may demonstrate inconsistency across different situations and contexts, they may also show instability over time. For instance, students' challenges could fluctuate between the initial weeks of a university term and the concluding weeks of the academic semester due to distinct characteristics and demands of tasks students face during these time blocks. Second, recall bias might be minimized, as specifying a short, recent time frame can make the recollection of relevant experiences more concrete and manageable for respondents. In the sections that follow, each of the subscales will be briefly discussed to provide an overview of their respective focuses and how they contribute to the overall understanding of academic challenges.

Motivation Challenge. The Motivation challenge, as defined by the 3 items: struggled with “believing I can do my work”, “feeling like my work was worth doing”, and “persisting when things got tough”, encompasses three aspects of students’ motivational challenges. This factor captures challenges related to the belief in one's own abilities in doing academic work (self-efficacy), the perceived value and meaningfulness of the work (value), and the capacity to continue efforts in the face of challenges (persistence). Collectively, these elements reflect the motivational underpinnings crucial for successful and sustained engagement in educational activities (Eccles & Wigfield, 2020; Pintrich, 2004; Schunk et al., 2013; Wigfield & Eccles, 2000; Winne & Hadwin, 2008).

Initiating and Sustaining Engagement (ISE). The *initiating and sustaining engagement* scale comprises four items that measure the internal motivation to start and persist with tasks, as well as the ability to manage attention. This factor encompasses difficulties in sparking interest in schoolwork (“wanting to do my schoolwork”), challenges in overcoming procrastination (putting things off until the last minute), difficulties in handling external and internal distractions, and the ability to maintain focus and concentration (“managing distractions”; “*focusing or concentrating*”). ISE captures the challenges related to both the initial drive to engage in academic activities and the sustained attention needed to remain engaged and productive.

Goal and Time Management Scale (GTM). The *goal and time management scale* defined by 4 items measuring struggles with effective time and task management (“Managing my time, tasks, or goals; Organizing time and tasks”), with pursuing academic goals (“Keeping my commitments or goals”), and with adhering to deadlines (“Meeting deadlines”). GTM reflects a student's struggles with organizing and controlling their academic workload efficiently and consistently, which is crucial for successful academic performance and productivity (Wolters & Brady, 2020).

Cognitive Challenges Scale (COG). The *cognitive challenge* scale, encompassing 6 items measuring struggles with cognitive processes involved in learning and understanding. This factor includes difficulties with recalling learned materials (“remembering things that I learned”), comprehend and internalize information (“understanding learning materials and concepts”; “identifying important ideas”), analyze and synthesize new concepts (“elaborating and connecting what I learned”; “applying my learning”), and communicate effectively (“Language

and communication”). Overall, the *COG* scale reflects struggles with the cognitive demands placed on students in processing, retaining, and applying academic content.

Metacognitive Scale (MET). The *metacognitive challenges* factor includes 8 items measuring challenges related to the ability to plan, monitor, and adjust one's approach to learning. It reflects struggles with the higher order thinking required to understand tasks and facilitate task engagement (“breaking tasks down”, “interpreting tasks and expectations”, “prioritizing or figuring out what is important”), set appropriate learning goals (“setting goals for studying”), monitor learning (“monitoring my task or learning progress”), and apply suitable strategies to optimize learning (“knowing how to tackle tasks”; “knowing how or when to fix strategies/study skills”; “applying strategies/techniques properly”). Metacognitive regulation, which includes metacognitive knowledge, awareness, monitoring, and control, is fundamental to self-regulated learning (Panadero, 2017; Winne, Winne, 2011; Winne & Hadwin, 1998).

Social and Emotional Scale (SE). The *social and emotional challenges* scale, comprising 6 items measuring struggles with interpersonal and emotional aspects of the student experience, including their ability to form connections and maintain relationships (“feeling connected”, “feeling lonely”, “managing relationships”), to find joy in their academic environment (“finding enjoyment at university”) and managing emotional wellbeing (“managing my emotions/feelings”; “taking care of my mental health and well-being”). Along with the relevant literature (e.g., Beaumont, et al., 2023; Davis & Hadwin, 2021; Kapil et. al, 2024; Pekrun & Linnenbrink-Garcia, 2012), the SE challenges highlight the significance of emotional wellbeing and sense of belonging in overall academic success and personal development. This subscale was added during the pandemic, a period characterized by limited interpersonal interactions due to restrictions and the widespread adoption of online instruction modes.

Awareness of the broad spectrum of challenges, including emotional and social aspects, supports the shift towards holistic educational models that simultaneously foster well-being and learning skills (White, 2023).

Criterion Measures

GPA Expectation (GPA-Exp). *GPA expectation*, measured by a single item, assesses students' anticipations regarding their academic performance for the current semester or term. This item asks participants to indicate the overall Grade Point Average (GPA) they realistically expect to achieve. The response options were structured on a scale to capture a range of expectations, from highest to lowest: High GPA, Strong GPA, Satisfactory GPA, Low GPA, and Failing GPA. These descriptors each accompanied by a quantified GPA range to illustrate what the descriptors meant at each site in terms of actual grades (e.g., Strong GPA (2.7 - 3.7)).

GPA Self-Efficacy (GPA-SE). *GPA Self-Efficacy*, measured by a single item, is designed to gauge students' confidence levels regarding their ability to achieve their expected GPA. After students specified their expected GPA using the provided descriptors and quantified ranges, they were asked to assess their confidence in attaining that GPA. The response options for this item range from *not at all confident* to *extremely confident*, with *neutral* serving as the midpoint. These measure students' self-efficacy and optimism towards their academic performance for the academic semester in which students were enrolled. Together with *GPA expectation*, this construct helps to capture the psychological aspect of students' academic expectations, adding depth to the understanding of their academic self-perception.

End of Semester GPA. End-of-semester GPAs for the students were obtained directly from the registrar offices of each institution. As each university employs its own GPA measurement scale, it was necessary to convert the GPAs to a common scale. This conversion

process entailed aligning the unique grading systems of each university to a standardized 5-point letter grade GPA scale. Numerical grades were converted to their equivalent letter grades based on conversion guides supplied by each institution.

The choice to employ a 5-point letter grade scale was influenced by its status as the most detailed scale used by one of the participating universities. Standardizing the GPAs on a common scale ensured consistency and facilitated meaningful comparisons across the different universities and their respective grading systems. On this converted scale, an "A" grade signifies "Excellent," a "B" grade indicates "Very Good," a "C" grade represents "Satisfactory," a "D" grade signifies "Barely Passed," and an "E" grade indicates "Failed."

Analytic Strategy

Item Response Theory (IRT)

Item Response Theory (IRT) is a psychometric method for evaluating items and properties of response scales. At its core, IRT evaluates individual items on a scale and aggregates this information to assess several overall properties of the scale. The foundational principle of IRT is that the likelihood of a particular response, such as selecting an option on a Likert scale, depends on an underlying latent attribute (De Ayala, 2013). This latent dimension, often referred to as *theta* (θ), is a person-centred parameter that indicates the extent to which an individual possesses a specific latent attribute. For instance, in the context of academic challenges, this parameter would measure a student's perceived level of a particular challenge. *Theta* is typically standardized to follow a standard normal distribution with an average of zero. In IRT, the focus is on modeling the relationship between person parameters (θ) and an individual's responses to items. Put differently, IRT is an item-level assessment of the measurement precision of scales across the range of the latent attribute (θ).

In addition to person parameters, there are also key item parameters in IRT, namely *discrimination* (α) and *threshold* (β) parameters describing properties of an item as a measure of the latent attribute. The term *discrimination* describes the ability of an item to differentiate between individuals who have different levels of the underlying attribute that the scale is intended to measure. On the other hand, *threshold*³ refers to the levels of the latent attribute above which individuals are likely to endorse higher response categories (Embretson, & Reise, 2013).

Through IRT, it is possible to identify items that offer less precise measurement and evaluate the reliability of each item in measuring the latent dimensions. The concept of *information* in IRT is a function of both discrimination and threshold parameters. Examining information parameters allow for estimating how much can be learned about a person's level on the latent attribute from their response to an item or a set of items within a scale. Such an examination provides insights into the degree to which items measure the intended range of the latent attribute and into the specific levels of latent attributes where they provide suboptimal measurement (Baker, 2001).

Graded Response Model

The self-reported academic challenge is measured using a 5-point ordinal Likert scale, therefore, Samejima's Graded Response Model (GRM; 1997, 2016), a polytomous variant of IRT, is used in this study. Discrimination or the *alpha* parameter in GRM indicates how well the items discriminate between different levels of the latent attribute. α can be interpreted as the

³ The term "difficulty" is commonly used to denote "threshold" in assessment contexts. However, we prefer the term "thresholds" as it more accurately reflects measures evaluating non-ability constructs.

strength of association between an item and the latent dimension, akin to a factor loading (De Ayala, 2013).

The discrimination parameter, can vary across a wide range, typically falling between zero to +2.5. According to Baker (2001), discrimination values between 0.01 to 0.34 can be regarded as very low, between 0.34 to 0.64 as low, between 0.65 to 1.34 as moderate, between 1.35 to 1.69 as high, and above 1.7 as very high.

Items with low discrimination are problematic because they imply that the likelihood of choosing a higher response option does not sufficiently increase as the latent attribute increases, which is undesired in most assessment contexts. Therefore, such items might not effectively measure the intended latent attribute and could potentially distort the overall scale's reliability.

In this study, GRM models with both constrained and unconstrained discrimination parameters were tested. In constrained models, the discrimination parameters across all items within a scale are set to be equal. Conversely, in unconstrained models, these discrimination parameters are freely estimated.

Also, In GRM, each response option is characterized by its unique threshold parameter or *beta*. The threshold for a response option refers to the level of the underlying dimension (e.g., academic challenges) at which there is an equal probability of selecting the next higher response option.

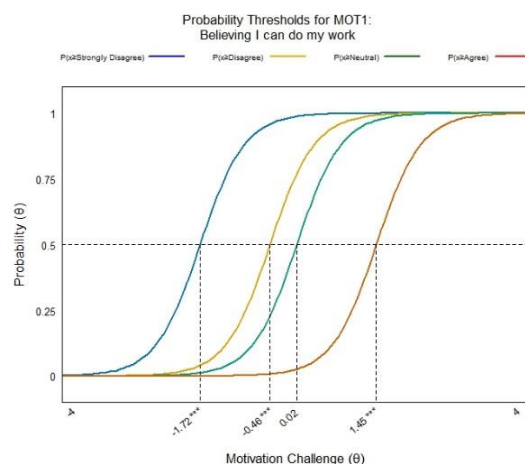
Graphs Depicting IRT Parameters

One of the key advantages of using unidimensional IRT is its ability to visually represent the quality of measurement in terms of how item and person parameters are related. To this end, Boundary Characteristic Curves (BCCs, Figure 1) and Item Response Category Characteristic Curves (ICCs, Figure, 2) provide a graphical representation of how the probability of endorsing a

particular response category changes across different levels of the latent attribute (Ostini & Nering, 2006). In addition, Item Information Curves (IICs) for each item (Figure 3a) and for the whole subscale (Figure 3b) show the amount of information that each item (or the subscale as a whole) provides at different levels of the latent attribute (Van Der Ark, 2001). IICs can be converted to reliability graphs indicating the reliability distribution for items and the subscales across the range of the latent attribute (Figure 4). The Expected Total Score plot (Figure 5) helps convert the latent attribute scores into the total score in the original observed units of the scale.

Figure 1

Boundary Characteristic Curves (BCCs) for an item of the Motivation Challenges Scale



Note. BCCs for the item *Believing I can do my work*. BCCs provide a visual representation of the probability thresholds' distances between response categories and the likelihood of endorsing each response category at different levels of the latent attribute (Motivation Challenge).

Color-coded curves response categories (e.g., (Strongly Disagree, Disagree, Neutral, Agree).

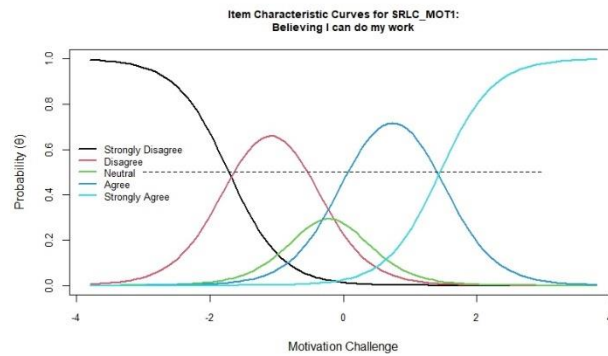
Vertical Axis (y-axis) represents the probability of endorsing each response category.

Horizontal Axis (x-axis) represents the levels of latent attribute (Motivation Challenge).

Numbers on the x-axis represent the threshold points where the probability of endorsing the next category becomes higher.

Figure 2

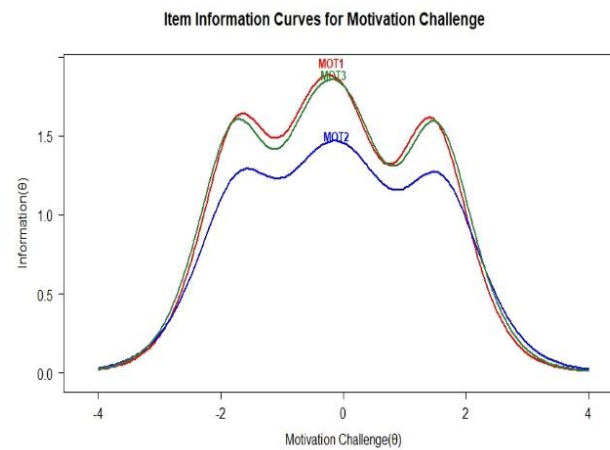
Category Characteristic Curves (CCCs) for an item of the Motivation Challenges Scale



Note. Each curve shows the probability of selecting a particular response category (e.g., Strongly Disagree, Disagree, etc.) across the levels of Motivation Challenge for an item. The x-axis represents the latent attribute, and the y-axis represents the probability. The points where two adjacent curves intersect reflect the thresholds where transition from one response option to the next is equally likely.

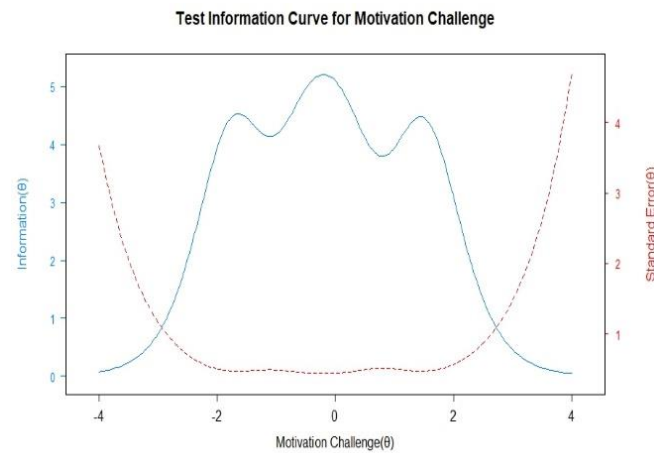
Figure 3

(a) Item Information Function (IIF) Graphs



Note. IIF displays the item information curves for the Motivation Challenges items. Each curve represents how much information an item provides about the latent attribute (Motivation Challenge) at different levels of it. IIF helps to identify at which levels of the latent attribute each item is most informative and is useful for selecting items that provide the most information at different levels of the latent attribute.

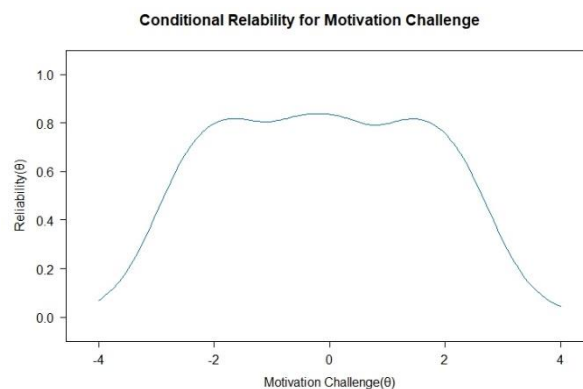
(b) Test Information Curve (TIC) and Standard Error Graphs



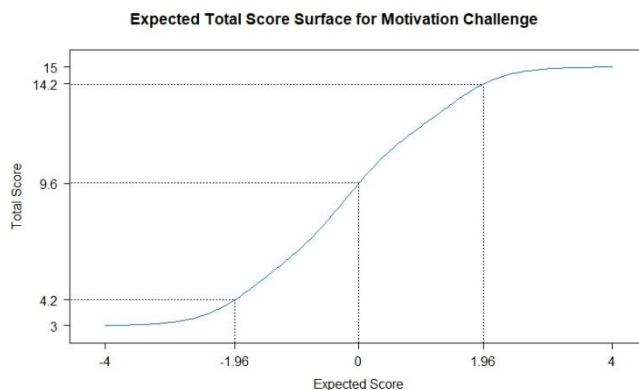
Note. TIC displays the test information curve and the standard error for the Motivation Challenges Scale. TIC shows the overall informativeness of the scale across different levels of the latent attribute. The test information curve aggregates the information from all items in the scale, while the standard error curve (inversely related to the information curve) shows the precision of measurement. Lower standard error means higher precision.

Figure 4

Conditional Reliability Graph



Note. This graph displays the conditional reliability of the Motivation Challenges Scale. It shows how reliability varies across different levels of the latent attribute and helps in identifying ranges of the latent attribute where the scale may need improvement to increase reliability. It shows the reliability of the scale at different levels of the latent attribute. Higher reliability values indicate more consistent and dependable measurement. Dotted vertical lines at specific attribute levels ($\theta = -2, 0, 2$) provide reference points for interpreting reliability at key points along the latent continuum.

Figure 5*Expected Total Score Graph*

Note. This graph displays the expected total score on the Motivation Challenges Scale as a function of the latent attribute (Motivation Challenge). It assists in interpreting what different latent scores mean in terms of the total scores in the original metrics of the scale.

Assumptions of IRT

Several fundamental assumptions must be satisfied in IRT, for the model to be valid and reliable. These include unidimensionality, local independence, monotonicity, and differential item functioning (Baker & Kim, 2017).

Unidimensionality. The assumption of unidimensionality posits all items on a scale should measure a single latent dimension. In other words, a single dimension should be sufficient to explain the response patterns to the items. This means the scale or test is measuring one, and only one, underlying attribute. If a test is truly unidimensional, a person's response to any one item should be predictable based on their responses to any one or set of the other items. If a test or scale is not unidimensional, the estimates of item parameters and person attributes may be biased or inaccurate, potentially leading to misleading conclusions about the measured trait, ability, or attribute. This assumption can be assessed using factor analysis, where ideally, all items should load significantly on a single factor.

Local Independence. The assumption of local independence implies that the latent attribute being measured should account for substantial portion of the variance in responses to the items within a scale. To test for local independence, the residual correlations between items loaded on a latent factor are examined. For a scale to be locally independent, these residual correlations should ideally be weak and not statistically significant, indicating that once the latent attribute is accounted for, the responses to different items within a scale are unrelated to each other (Edwards, et al., 2017).

Monotonicity. Monotonicity in IRT refers to an important assumption about the relationship between the estimated latent dimension being measured (e.g., trait, ability, or attribute) and the probability of a given response to an item. Specifically, the monotonicity means that as the level of the latent attribute increases, the probability of selecting a higher category response (or demonstrating a higher level of ability/performance on an item) should also increase, or at least not decrease (Kang, et al., 2018). In IRT analysis, checking for monotonicity is an important step in ensuring that the items on a scale or test are functioning properly and are appropriate for measuring the intended latent dimension. Violation of this assumption indicates a problem with the item itself or with the way the latent dimension is being conceptualized or operationalized. An item that does not exhibit monotonicity might be confusing, poorly worded, or not aligned properly with the underlying construct, leading to responses that do not reflect the intended measurement of the latent attribute.

Differential Item Functioning. Assumption of Differential Item Functioning (DIF) ensures that the item response patterns are consistent across different demographic groups, such as gender. The absence of DIF suggests that the items function similarly across these groups,

implying that the scale is fair and unbiased in measuring the latent attribute across the diversity of population it intends to represent (Chalmers et al., 2016).

Each of these assumptions plays a crucial role in ensuring the integrity and applicability of an IRT model, making them critical requirements in the development and evaluation of psychometric scales.

Results

Checking the Assumptions

Descriptive statistics can be found in the supplementary materials B. Considering that, on average, less than 3 percent of the academic challenges responses were missing, pairwise deletion was employed in analyses that necessitated prior management of missing data. The skewness of the items ranged from -1.07 to 0.33, indicating that while most items were fairly symmetric, some exhibited a slight negative skew. Similarly, the kurtosis values ranged from -1.21 to 0.77, suggesting that the items generally had light tails with slight deviations from the normal distribution. These values indicated that while the data show some departures from “perfect” normality, the deviations are within a range that is considered acceptable for many statistical analyses including factor analysis (Byrne, 2010; Hair et al., 2010; Kline, 2015, Tabachnick & Fidell, 2013).

The data analysis was conducted using R packages: Mirt (v1.41; Chalmers, 2012) for Item Response Theory (IRT) analyses, Lavaan (v0.6-16; Rosseel, 2012) for Confirmatory Factor Analyses (CFAs) and computing regression factor scores (Thurstone, 1934), Psych (v2.4.1; Revelle, 2015, 2020, 2024) for Exploratory Factor Analysis (EFA), Polycor (v0.8-1; Fox & Dusa, 2022) to obtain polychoric and polyserial correlation coefficients (Dragow, 1988), and mokken (v2.4; van der Ark, 2007, 2010, 2012) for checking the monotonicity assumption.

Additionally, results obtained from Mirt were cross-verified with the ltm package (v1.2-0; Rizopoulos, 2006).

Unidimensionality. The initial phase of the study involved conducting exploratory and confirmatory factor analyses to investigate the unidimensionality of the *academic challenge* subscales. Here, unidimensionality implies the covariance among the items within each subscale is exclusively accounted for by a single-factor structure.

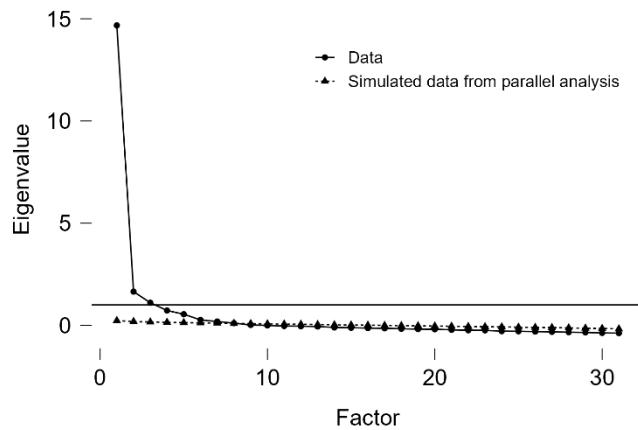
Eigenvalues and scree plots derived from exploratory factor analysis (EFA) were used to examine the unidimensionality assumption for each of the subscales (see supplementary materials C). The eigenvalues for the first factors were much higher than the eigenvalues for the second factors across all the subscales, indicating a strong single factor structure. For instance, the eigenvalues for the first factors were 1.85 for MOT, 2.11 for ISE, 2.5 for GTM, 3.8 for COG, 5.5 for MET, and 3.24 for SE, with the highest second-dimension eigenvalues being 0.31. The lowest ratio of the first to the second eigenvalue across subscales was 12.6, further supporting the unidimensionality of scales. Comparing these eigenvalues with those obtained from random simulated samples with the same number of variables and observations, a method known as parallel analysis (Horn, 1965), supported that subscales can be explained by a single factor. The proportion of variance explained by the single-factor measurement models ranged from 0.53 to 0.68 across the scales (see supplementary materials C).

For the general academic challenges scale, hereafter referred to as AC-g, the scree plot shown in Figure 6 suggested the extraction of three factors. However, the eigenvalue for the first factor was 14.7, and the ratio of the eigenvalue of the first to the second factor was 8.9. The first factor accounted for about 48% of the variability in the data, while the second and third factors

accounted for about 6% and 4% of the response variability, respectively. This observation suggested a single-factor structure represented about half the variability across items.

Figure 6

EFA Scree Plot For the AC-g Scale



Following the EFA, confirmatory factor analysis (CFA) was conducted to further estimate the global fit properties of the unidimensional scales. Given that the response scale in the study was a 5-point ordinal Likert scale, the Diagonally Weighted Least Squares estimator (DWLS, see Mîndril, 2010) with robust standard errors was chosen for the analysis.

As shown in Table 1, the chi-square fit statistic for all scales was significant. However, as Kline (2015) notes, this statistic is sensitive to sample size. This means that in larger samples, the chi-square value tends to be higher, and the likelihood of it being significant increases.

Table 1
Global Fit Indices for the Scales

Factor	X^2	df	p	CFI	TLI	RMSEA	RMSEA 90% Lower CI	RMSEA 90% Upper CI	SRMR
MOT	0	0	NA	1	1	0	0	0	0
ISE	35.328	2	0	0.996	0.989	0.045	0.033	0.058	0.023
GTM	130.794	2	0	0.992	0.975	0.085	0.073	0.098	0.039
COG	526.662	9	0	0.989	0.981	0.075	0.07	0.081	0.053
MET	617.729	20	0	0.994	0.991	0.054	0.05	0.057	0.043
SE	270.279	9	0	0.99	0.984	0.06	0.054	0.066	0.043
AC-g	10199.07	434	0	0.97	0.968	0.072	0.07	0.073	0.073

The fit indices were generally satisfactory, with the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) exceeding 0.95, the Root Mean Square Error of Approximation (RMSEA) 90% upper interval being below 0.08, and the Standardized Root Mean Square Residual (SRMR) being below 0.05 for the subscales and below 0.08 for the AC-g. The only exception was the RMSEA for GTM being slightly higher at 0.085 [90% CI: 0.073-0.098]. Modification indices suggested that allowing the residual covariance between items GTM2 and GTM4 to covary would lead to a significant improvement in model fit, as indicated by the RMSEA reducing to 0.025 with a 90% confidence interval ranging from 0.013 to 0.041. The need to estimate the residual covariance between these two items and the resulting improvement in model fit can be interpreted as evidence of the presence of method effects, specifically due to similar wordings in the items. In this instance, the similar phrasings of GTM2 (*managing my time, tasks, or goals*) and GTM4 (*organizing time and tasks*) could have prompted respondents to answer these items similarly jointly due to their thematic congruence and overlap in wording. Allowing the residual covariance of these items to covary adjusted for this method effect, leading

to a more accurate representation of the underlying construct. Also, the fit indices for the *motivation* challenge scale were not estimated due to it having only three items and being just-identified (see Hoyle, 2012; Kline, 2015).

All factor loadings were significant at $p < 0.001$, ranging from standardized estimates of 0.71 to 0.76 for MOT, 0.46 to 0.8 for ISE, 0.62 to 0.83 for GTM, 0.66 to 0.85 for COG, 0.72 to 0.83 for MET, 0.53 to 0.82 for SE, and 0.38 to 0.78 for the AC-g (see supplementary materials D & E).

The internal consistency reliability was examined by computing McDonald's Omega (ω , McDonald, 1999) for each scale. The Omega values were 0.78 for MOT, 0.76 for ISE, 0.83 for GTM, 0.89 for COG, 0.93 for MET, 0.84 for SE, and 0.96 for the AC-g, indicating a range of adequate to high reliability across the scales (see supplementary materials F).

When analyzing the item reliability statistics, it was noted that eliminating certain items from their respective scales does not adversely affect the scales' reliability. Specifically, the exclusion of item GTM3, SE2, and ISE2 did not diminish the reliability of the GTM, SE, and ISE subscales. In fact, removing ISE2 even marginally enhanced the point estimate of reliability of the ISE scale (see supplementary materials G). Similarly, the removal of ISE2, ISE3, and SE2 had no detrimental impact on the internal consistency of AC-g. Further investigation into the item-rest correlations for these items revealed a moderate correlation with the total score of their respective scales (average $r = .46$), indicating that these items are comparatively less aligned with their intended constructs (see supplementary materials H).

The findings from both exploratory and confirmatory factor analyses, alongside internal consistency estimates, provided sufficient evidence to support the unidimensionality of each of the six subscales and the AC-g scale.

Local Independence. To assess the local independence of items, examining the correlation of residuals is a common practice. Residuals represent the discrepancies between the scores predicted by the model and the actual observed scores. If residuals between item pairs remain correlated after controlling for the latent factor, this suggests that the items are related due to factors other than the common latent attribute, indicating local dependence. Such dependence can be caused by various factors, including unmodeled latent variables, method effects, or response dependencies resulting from the sequencing of items (Andrich & Marais, 2019).

In this study, Yen's (1993) index of local dependence (Q_3) was used to evaluate the assumption of local independence. A Q_3 value above .3 was chosen as indicative of local dependence between item pairs (Marais, 2013). None of the estimates exceeded this threshold (see supplementary materials I). The highest Q_3 estimates was for the pair consisting of items Cog3 and Cog4, with a Q_3 value of .26. Item Cog3 captures students' struggles with language and communication, whereas Cog4 pertains to challenges in identifying important ideas. This value suggested a shared variance of less than 7% for the conditional correlation coefficients between these two items. Given the additional evidence supporting the unidimensionality of the scales, these findings supported that the items within each subscale are locally independent.

Evaluating the local independence among the 31 items of the AC-g scale was done with an awareness that we were intentionally ignoring specific academic challenge factors to achieve a unidimensional factor for practical application. Nine pairs of items were identified that exceeded this .3 cut-off (see supplementary materials J). Notably, no local dependence was observed across items from different factors comprising the academic challenges scale. Local dependencies were solely observed in items from the same factor. Among these, the pairs Cog1-

Cog2, ISE3-ISE4, and SE4-SE5 shared 29%, 22%, and 15% of the variance in their residuals, respectively, indicating that 71%, 78%, and 85% of their residual variances were unique. However, while some degrees of dependencies were observed due to multidimensionality, responses to each item were substantially more influenced by the latent attribute than by any potential unmodeled factors. As residuals of items within scales were found to be correlated, such potential unmodeled factors were likely be the specific factors that were excluded in the current unidimensional analysis. For example, the variance accounted for by items ISE3-ISE4, beyond the general academic challenges construct, pertains to the regulation of attention which was measured by the specific factor of *initiating and sustaining engagement* in the correlated factor model.

Despite such indications of local dependence, given the sufficient evidence supporting unidimensionality, these results were not deemed sufficiently problematic to hinder further investigation of the AC-g scale. The literature suggests that the presence of a dominant general factor in the data means that multidimensionality has minimal impact on IRT item parameter estimates and the associated person parameters (Gibbons, et al., 2007; Ip, 2010).

That being said, graded response models with and without an item from each pair of locally dependent items were compared as part of the main analysis. This comparison aimed to investigate the potential bias in IRT parameters introduced by such local dependencies (see Edward, et al., 2017). To elaborate, failing to account for local dependencies (LD) in IRT models is conceptually similar to the under-factoring in exploratory factor analysis (Ip, 2010). In the context of the current study, instead of using a multidimensional or bifactor polytomous IRT, the six factors of academic challenges scales were first evaluated as unidimensional constructs and then forced to load on a single dimension representing a general academic challenges scale. This

strategy was used to evaluate the measurement precision of items in measuring unidimensional constructs that can be used independently in research and practice. If a model only estimates one factor when there are actually more (with one possibly being a minor method effect involving just a couple of items), the resulting unidimensional factor ends up combining the intended and the method effect factors to maximize variance explained. Yen (1984) and others have observed that in such cases, especially in short tests, local dependencies can cause the model to estimate a latent dimension that deviates from the intended one, favoring a dimension that better accounts for the high covariance between the locally dependent items. This results in the locally dependent items having a disproportionately strong relationship with this "compromise" latent dimension, indicated by high slope parameter values. On the other hand, non-locally dependent items relate more weakly to this compromised factor than they would to the true latent dimension (Edwards, et al., 2017).

Removing one item from a locally dependent pair eliminates the model's tendency to deviate from the true latent dimension, allowing for a more accurate alignment with the true latent dimension. For this reason, an item from each of the 9 pairs identified as locally dependent (with Q_3 residuals exceeding 0.3) was excluded from the analysis: GTM4, ISE3, SE3, COG2, COG5, COG3, SE5, SE6, and SE2. After their removal, the residual matrix of Q_3 was revisited to ensure no residual covariances surpassed the .3 cut-off. Subsequently, the model was re-estimated. A comparison was made of discrimination and threshold parameters between the original model, which included all items, and the revised model, from which the locally dependent items were removed. The results revealed that the IRT estimates from both models were remarkably similar. This similarity suggested that local dependencies did not importantly

affect the model's ability to estimate the true AC-g, further affirming the stability of the IRT parameters even when the multidimensionality of factors was not considered.

Monotonicity. The concept of monotonicity in IRT is another assumption for understanding how respondents' true scores on a latent attribute (θ) relate to their responses on a Likert scale. The assumption is that as the level of the latent attribute increases, the probability of selecting a higher point on a Likert scale should also increase. In this study, monotonicity was tested using Mokken scaling, a non-parametric form of item response theory that does not assume a specific distributional form for item responses (Mokken, 1971; Sijtsma & van der Ark, 2017).

Monotonicity plots (e.g., Figure 7) were used to visually examine if this assumption is met. In these plots, the x-axis represents the overall sum score on the scale excluding the score from the item under investigation. The y-axis represents the item response function (IRF) for individual items. These plots visualize the probability of a particular response to an item against the level of the latent attribute. Following the individual item response function plots, we also generated a response function that aggregates all items within a scale, providing a holistic view of how responses vary within a scale. In an ideal scenario, the IRFs for all items should show a relatively diagonal line, indicating a direct and consistent relationship between the scale's total score and the item responses. What should be observed to meet the monotonicity assumption is the absence of any downward dips in the IRFs. Such dips would suggest that higher levels of the latent attribute are associated with lower responses on the ordered Likert scale, which is counterintuitive in this context.

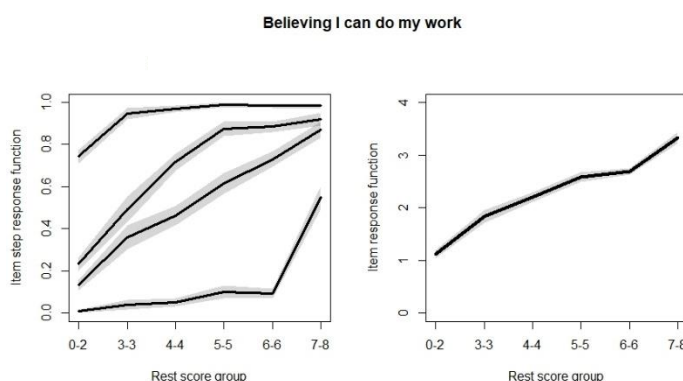
The Mokken analysis (see supplementary materials K) demonstrated that only ISE1 and SE2 among the items of the subscales were non-monotone. However, the IRF plots

(supplementary materials L) demonstrated that the item response functions are generally monotone for all the items of the subscales.

Evaluating the monotonicity assumption for the AC-g scale supported that the item response functions are also generally monotonic; however, some degree of non-monotonicity was observed, particularly for items ISE2 and SE2, with a milder degree observed in items ISE3 and SE6 (see supplementary materials M, supplementary materials N). Given that the overall pattern of the item response functions only slightly deviated a monotonic shape, we concluded

Figure 7

Monotonicity Plot for an Exemplar Item



Note. Monotonicity Plots for all items can be found in the supplementary materials L & M

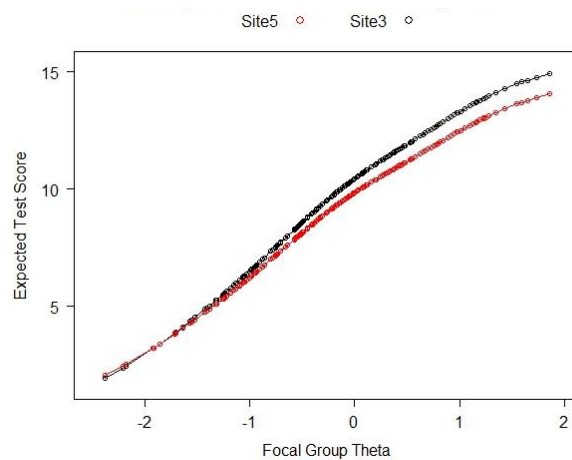
that these items could be retained within the scale at this stage. Nonetheless, the reasons behind their slight unexpected behavior, along with decisions regarding whether to keep, modify, or remove them from the scale, was deferred until after the evaluation of the IRT model parameters.

Differential Item Functioning. As the sample comprised students from five universities across Canada. To explore potential differential item functioning (DIF) across these sites, various indices were calculated as outlined by Meade (2010): SIDS (Signed Item Difference in Sample), UIDS (Unsigned Item Difference in Sample), SIDN (Signed Item Difference in Normal Distribution), UIDN (Unsigned Item Difference in Normal Distribution), and ESSD (Expected Score Standardized Difference) (see supplementary materials O & Q).

The analysis involved comparing signed and unsigned estimates, where discrepancies between these estimates could suggest cancelation in DIF. Such cancelation implies that DIF might be nonuniform, with the focal group showing lower expected scores than the reference group in certain intervals of the latent attribute, but higher in others. This variability could lead to small DF estimates when averaged across all respondents within a single item. In this study, the proximity of signed and unsigned values indicated a lack of nonuniform DIF across sites.

We generated expected score plots (e.g., Figure 8) to visually assess the presence and nature of DIF for items where SIDS values exceeded .3 (Marais, 2013). This cut-off was selected as it was one quarter of the average standard deviation of the items. A SIDS value greater than .3 indicates that, on average, members of the focal group in the current sample would score .3 points more (+.3) or less (-.3) on an item compared to members of the reference group with similar estimated latent attribute scores. Analysis of these plots revealed that the relationships between latent attribute levels and expected scores was relatively consistent in *shape* across different sites, with the observed differences being slight variations in the *levels* of expected scores (see, supplementary materials P & R). Since the goal of collecting data from multiple sites was to obtain a representative sample of students, it was concluded there was no concern regarding DIF and that combining site-specific differences would yield an appropriate representation of the population.

Figure 8
Differential Item Functioning Plot for an Item (ISE2) across Two Participating Sites



Note. DIF Plots for items across the participating sites can be found in supplementary materials P & R

Item Fit

The Orlando-Thissen-Bjorner item fit statistic, $S-X^2$ (Orlando & Thissen, 2000, 2003), was employed to assess the absolute fit of the model to each item. In the context of $S-X^2$, a non-significant value typically indicates a good fit of the model to the data. However, in the present study, the $S-X^2$ estimates for most items were lower than a p -value of .005 (see supplementary materials S & T). These estimates did not undergo substantial changes even after applying the Benjamini-Hochberg p -value adjustment (Benjamini & Hochberg, 1995). It is important to recognize that $S-X^2$ is sensitive to factors such as sample size, test length, and the occurrence of multiple comparisons. These factors (i.e., larger sample, shorter tests, multiple comparisons) can inflate the estimate, thereby increasing the likelihood of a Type I error and yielding significant values (Kang & Chen, 2008).

To determine whether items were erroneously identified as misfits, RMSEA (Root Mean Square Error of Approximation) estimates based on the $S-X^2$ values (termed RMSEA $S-X^2$) were examined. Generally, RMSEA values of .06 or lower are considered indicative of an acceptable

fit (Maydeu-Olivares, 2013). In this study, all RMSEA values were below .05, suggesting adequate item-level model-data fit.

Constrained vs. Unconstrained GRM Models

The Graded Response Model (GRM) is categorized into two fundamental types, constrained and unconstrained. The constrained model assumes equal discrimination parameters across all items, suggesting that every item is equally effective at discriminating among respondents across different levels of the latent attribute. Conversely, the unconstrained model recognizes that each item may have unique discrimination parameters, allowing for variability in how effectively each item discriminates among respondents' latent attribute levels.

In this study, I performed a comparative analysis of these two model types. For the subscales, likelihood ratio tests, alongside AIC and BIC values, revealed that, except for one scale (i.e., motivation), the unconstrained models—which allow for the free estimation of discrimination parameters for each item—consistently yielded a superior fit to the data. For the AC-g, the findings also indicated a significantly better fit for the unconstrained model compared to the constrained model. Consequently, in all subsequent analyses, we opted not to impose any constraints on the discrimination parameters and allowed for estimation of unique discrimination values for each item.

The results of the IRT analyses for the six scales of academic challenges, as well as for the AC-g, will be presented separately. This approach is due to the detailed nature of the analyses and the necessity for systematic reporting. As the study comprises 7 discrete scales, reporting results for each scale separately allows readers to only focus on the results for a single scale if desired. Consequently, the structure of the reporting will remain consistent across scales with

general explanations repeated as needed. The first items from each scale were used to illustrate and guide the interpretation of the figures and tables presented within the results section.

IRT Analyses of the Motivation Challenge Scale (MOT)

In the initial phase of the analysis, frequency distribution tables for the motivation challenges items were constructed (Table 2). Comparisons of the frequency distribution of item responses are helpful because items with different distributions tend to have lower linear association with each other. The distribution frequencies for the three items of the *motivation challenge* scale did not vary much and were relatively consistent. On average, around 8% of respondents selected the lowest response option, *strongly disagree*, while about 10% chose the highest response option, *strongly agree*. Notably, the most frequently selected response was option 4, *agree*, with about an average of 37% of respondents choosing this option.

Table 2

The Frequency Distributions of the Motivation Challenges Items (N=2671)

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
MOT1					
Frequency	214	694	428	1034	299
Percent	8.02	26	16.04	38.74	11.2
Cumulative%	8.02	34.02	50.06	88.8	100
MOT2					
Frequency	246	740	467	934	280
Percent	9.22	27.75	17.51	35.02	10.5
Cumulative%	9.22	36.97	54.48	89.5	100
MOT3					
Frequency	200	713	475	1005	275
Percent	7.5	26.72	17.8	37.67	10.31
Cumulative%	7.5	34.22	52.02	89.69	100

Concerning item discrimination (*alpha*), all items exhibited high and relatively uniform parameters, ranging from 2.21 to 2.52 (Table 3). According to Baker and Kim (2017),

discrimination values above 1.7 could be interpreted as very high. High discrimination values indicate that the items are effective in differentiating between individuals with varying levels of the perceived motivation challenges. As for the estimated thresholds (*beta*), these values varied from -1.77 to 1.56 and were comparatively similar in terms of their magnitude across items (Table 3).

To illustrate the threshold parameters using the first item, struggling with *believing I can do my work* (Table 3), an average respondent with a latent score of -1.72 had a 50% probability of choosing response option *disagree* (2) or higher versus choosing lower response (i.e., *strongly disagree*). An average respondent with a latent score of -0.46 had an equal (50%) chance of selecting response option 3 (*neutral*) or higher as opposed to choosing a lower response category. An average respondent with a latent score of 0.02 had an equal (50%) chance of selecting response option 4 (*agree*) or higher versus choosing a lower response category. Finally, an individual with a latent score of 1.45 was equally likely to choose response 5 (*strongly agree*) as they were to choose any lower response, reflecting a 50% probability in either case. Interpreting the latent score as akin to a z-score, which is normally distributed across a population with a mean of zero, if an individual's latent score is above 1.45, they are significantly above the average level for the motivation challenge being measured. In the context of survey questions about their experiences over the past two weeks, such a person would be likely to select the highest response option, *strongly agree*, when asked about the extent to which they felt their work was worth doing (MOT1).

Table 3*Item Response Analysis of the Motivation Challenges Scale*

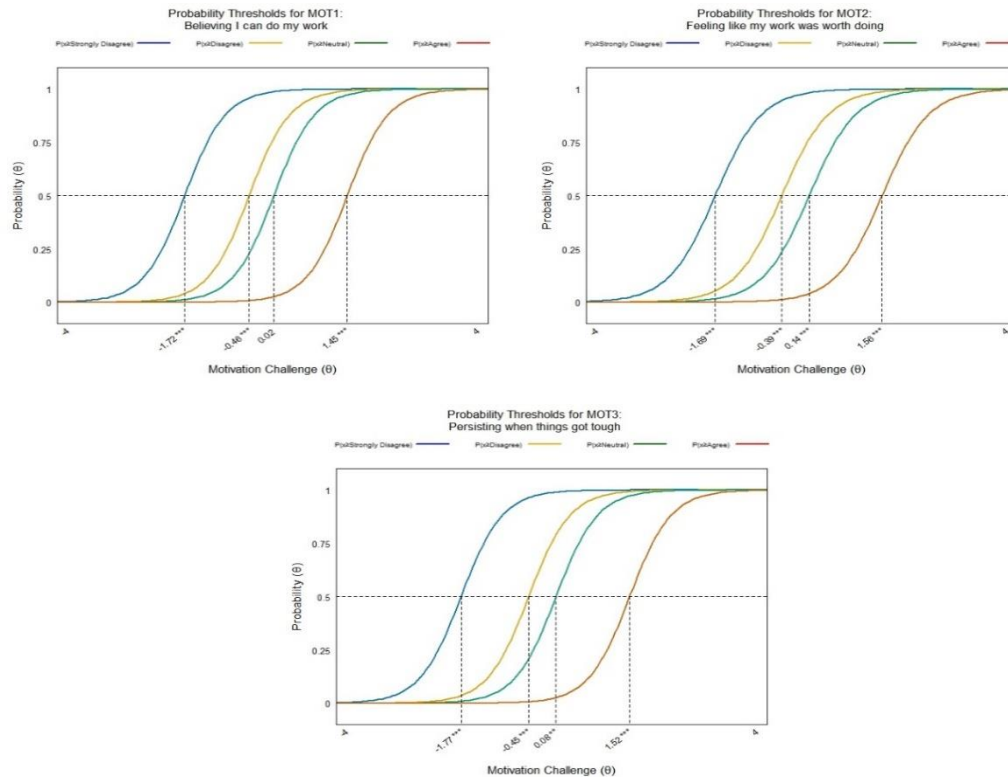
	Coef.	SE	Z	p	CI_2.5	CI_97.5
MOT1: Believing I can do my work						
Discrimination (α)	2.52	0.12	21.88	0.00	2.29	2.75
Thresholds (β)						
>2	-1.72	0.05	-31.41	0.00	-1.82	-1.61
>3	-0.46	0.07	-6.90	0.00	-0.59	-0.33
>4	0.02	0.05	0.40	0.69	-0.08	0.12
>5	1.45	0.55	2.62	0.01	0.36	2.54
MOT2: Feeling like my work was worth doing						
Discrimination (α)	2.21	0.09	23.60	0.00	2.03	2.39
Thresholds (β)						
>2	-1.69	0.06	-30.06	0.00	-1.80	-1.58
>3	-0.39	0.06	-6.80	0.00	-0.50	-0.28
>4	0.14	0.05	3.05	0.00	0.05	0.24
>5	1.56	0.50	3.12	0.00	0.58	2.54
MOT3: Persisting when things got tough						
Discrimination (α)	2.50	0.11	22.00	0.00	2.27	2.72
Thresholds (β)						
>2	-1.77	0.06	-31.56	0.00	-1.88	-1.66
>3	-0.45	0.07	-6.64	0.00	-0.58	-0.32
>4	0.08	0.05	1.56	0.12	-0.02	0.19
>5	1.52	0.64	2.36	0.02	0.25	2.78

Across the items, the threshold parameters (*beta*) were fairly consistent. For instance, for the response *strongly agree* the threshold values were 1.45, 1.56, and 1.52 for item1 to 3, respectively. However, within each item, the differences between threshold parameters for consecutive response categories varied (Figure 9, also see supplementary materials U). For instance, in the first item, the distance between the thresholds for *strongly disagree-disagree* (b1-b2) and *disagree-neutral* (b2-b3) was 1.26, while the distance between thresholds for *disagree-neutral* (b2-b3) and *neutral-agree* (b3-b4) was substantially smaller, at 0.48, and the distance between thresholds for *neutral-agree* (b3-b4) and *agree-strongly agree* (b4-b5) was 1.43. Consequently, an individual's total score on the MOT challenge scale might differ from scores

obtained using Classical Test Theory (CTT), where item scores are simply summed as they are assumed to be equally spaced (see Willse, 2022).

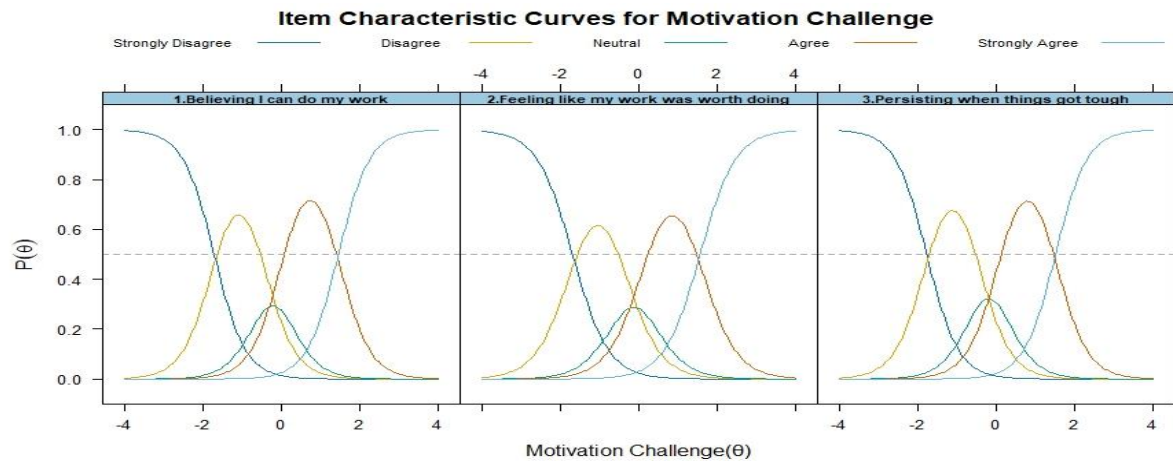
Figure 9

Boundary Characteristic Curves for each Item of the Motivation Challenges Scale



The Category Characteristic Curves (CCCs) in Figure 10 demonstrated the relationship between participants' locations on the motivation challenge latent attribute and the probability of choosing specific response options for each item. These curves indicate the thresholds at which a higher response option becomes more probable. The points where two adjacent curves intersect reflect the thresholds where transition from one response option to the next is equally likely. For example, for the first item, respondents with a latent score below -1.72 are more likely to respond with *strongly disagree*, while those with scores between -1.72 and -0.46 are most likely to choose *disagree*, and so forth. Additionally, within this score range, these curves indicated the probability of selecting other response options as well; for instance, the likelihood of selecting

Figure 10
Category Characteristic Curves for each Item of the Motivation Challenges Scale



option *strongly agree* are about zero in this range while the probabilities of selecting *agree* are non-zero.

Observing the latent score range where participants are inclined to select *neutral* as the midpoint of the Likert scale (a score between -0.46 and 0.2) showed that within this confined range—where the likelihood of choosing *neutral* was at its peak (e.g., approximately 30% for MOT1)—the probabilities of selecting response options *disagree* and *agree* exceeded that of *neutral* for item MOT1 and MOT2. This indicated lower measurement precision for the midpoint of the Likert scale, compared to the adjacent options.

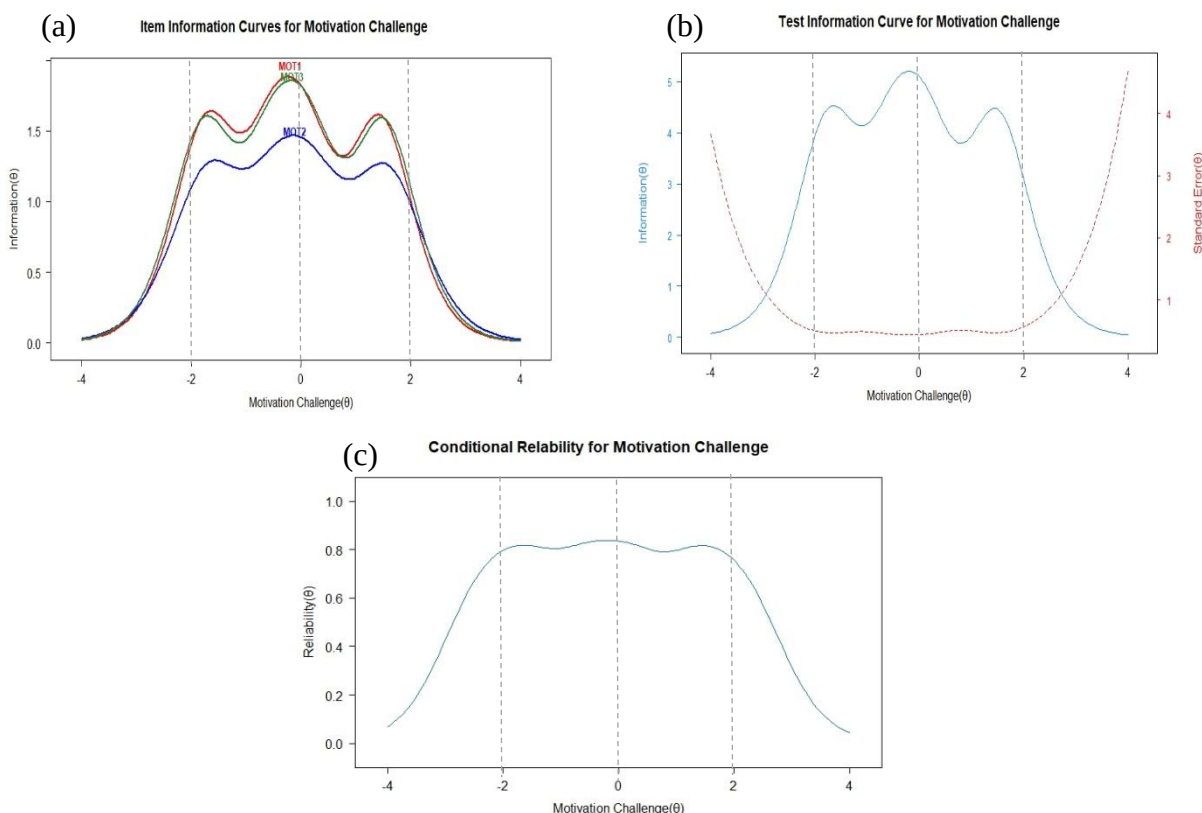
We also examined the item information function for each item (see Figure 11a). In IRT, the information function integrates the characteristics of threshold and discrimination parameters to indicate how much information each item contributes at different points along the latent continuum. Essentially, by investigating these plots, we can identify the level on the motivation challenge latent continuum where each item provides the greatest information or the highest measurement reliability. If a range on the latent continuum is underrepresented by the current items, adding a new item that targets this specific range could improve the overall measurement.

All items exhibited similar levels of information, with item MOT2 providing slightly less information compared to the other two items. Overall, these items reliably covered the distribution of the motivation challenge latent continuum (*theta*), covering a range from -2 to 1.9. Additionally, the plot indicated that we obtained reliable information at an average *theta* value of 0, with item MOT2 yielding an information value of approximately 1.3, and items 1 and 3 providing higher values, around 1.9.

The item information functions were aggregated to form the test information function, which was used to assess the precision of the entire scale at various levels of the motivation challenge latent attribute. A higher value in the test information function signifies higher precision in measurement. The standard error of measurement is represented by the red dotted line in the plots. Figure 11b, indicated that the three items collectively offered substantial information for measuring motivation challenges in participants scoring between -2 and 1.9. This observation indicated that the *motivation* challenge scale demonstrated high reliability and low measurement error within this range. The scale provided the most information and the lowest measurement error around average *theta* levels (information=5). However, the information level decreased for *theta* values below -2 and above 2. The reliability for varying levels of motivation challenge is depicted in Figure 11c. As shown, the reliability was approximately 0.8 for *theta* levels between -2 and 2, but it diminished as the *theta* values moved away from this range.

Figure 11

- (a) Item Information Function Graphs for the Motivation Challenges Items
 (b) Information and Standard Error Graphs for the Motivation Challenges Scale
 (c) Conditional Reliability for the Motivation Challenges Scale

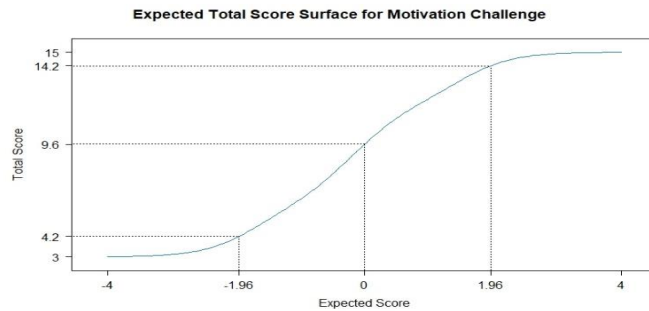


Note. Dotted vertical lines at $\theta = -2, 0, \text{ and } 2$.

Figure 12 displays the expected scores derived from the test characteristic curves, plotted against the latent continuum (*motivation challenges*). This plot can be used to convert *theta* values into corresponding total scores in the unit of the observed variables.

Converting latent attribute scores to sum scores helps in understanding the distribution and tendencies of responses within the MOT challenge context. Given that the *motivation* challenge scale comprises three items, each with a five-point Likert scale, the range of expected scores falls between 3 and 15. The results indicated that participants scoring below -1.96 on the latent continuum would have expected scores of 4.2 or lower, suggesting they are most likely to

Figure 12
Expected Total Score Graph for the Motivation Challenges Scale



select response options 1 or 2 on the Likert scale. By choosing ± 1.96 , which represents two standard deviations in the standard normal distribution, we can infer that approximately 95% of randomly selected participants would score between 4.2 and 14.2 with an average of 9.6.

IRT Analyses of the Initiating and Sustaining Engagement Challenges Scale (ISE)

Frequency distribution tables for all *initiating and sustaining engagement challenges* items were examined as the first step (Table 4). The varied frequencies across items like ISE2 struggling with *avoiding putting things off until the last minute (procrastinating)* and ISE1 struggling with *wanting to do my schoolwork* suggested distinct item characteristics. For example, 32.3% selected response option *agree* for ISE2, whereas a higher 43.5% responded similarly to ISE1. This variation is notable as items with different distributions tend to have lower linear correlations with each other. Regarding the lowest response option (*strongly disagree*), 5% selected this option for both ISE1 and ISE3, while a higher 9.1% chose it for ISE2. Conversely, for the highest response (*strongly agree*), about 20.5% selected this for ISE1, with ISE4 showing a higher percentage at 26%. Also, *agree* emerged as the most common response option across categories.

Table 4*The Frequency Distributions of the ISE Challenges Items (N=2671)*

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
ISE1					
Frequency	136	431	393	1161	549
Percent	5.09	16.14	14.72	43.48	20.56
Cumulative%	5.09	21.24	35.96	79.44	100
ISE2					
Frequency	244	488	398	861	678
Percent	9.14	18.28	14.91	32.26	25.4
Cumulative%	9.14	27.43	42.34	74.6	100
ISE3					
Frequency	134	431	444	1091	569
Percent	5.02	16.15	16.64	40.88	21.32
Cumulative%	5.02	21.17	37.8	78.68	100
ISE4					
Frequency	124	393	372	1080	698
Percent	4.65	14.74	13.95	40.49	26.17
Cumulative%	4.65	19.39	33.33	73.83	100

Examining item discrimination parameters (Table 5) revealed that, except for ISE2, all items showed very high discrimination values, ranging from 1.78 to 3.12, indicating their effectiveness in distinguishing between different levels of the ISE challenge. However, ISE2 showed moderate discrimination at 1.1. According to Baker and Kim (2017), values from 0.65 to 1.34 can be interpreted as moderate, and those above 1.7 as very high discrimination.

Table 5
Item Response Analysis of the ISE Challenges Scale

	Coef.	SE	Z	p	CI_2.5	CI_97.5
ISE1: Wanting to do my schoolwork						
Discrimination (α)	1.78	0.07	24.42	0.00	1.64	1.92
Thresholds (β)						
>2	-2.30	0.08	-28.05	0.00	-2.46	-2.14
>3	-1.10	0.09	-12.88	0.00	-1.27	-0.93
>4	-0.51	0.07	-7.21	0.00	-0.64	-0.37
>5	1.13	0.12	9.40	0.00	0.89	1.36
ISE2: Putting things off until the last minute (procrastinating)						
Discrimination (α)	1.10	0.05	20.81	0.00	1.00	1.21
Thresholds (β)						
>2	-2.47	0.11	-21.62	0.00	-2.69	-2.24
>3	-1.10	0.09	-12.52	0.00	-1.27	-0.92
>4	-0.38	0.07	-5.78	0.00	-0.51	-0.25
>5	1.17	0.11	10.47	0.00	0.95	1.39
ISE3: Managing distractions						
Discrimination (α)	2.50	0.11	23.27	0.00	2.29	2.71
Thresholds (β)						
>2	-2.00	0.06	-31.85	0.00	-2.12	-1.88
>3	-0.96	0.08	-11.47	0.00	-1.13	-0.80
>4	-0.39	0.07	-5.66	0.00	-0.53	-0.26
>5	0.95	0.13	7.19	0.00	0.69	1.21
ISE4: Focusing or concentrating						
Discrimination (α)	3.12	0.16	19.21	0.00	2.80	3.43
Thresholds (β)						
>2	-1.92	0.06	-33.06	0.00	-2.03	-1.80
>3	-0.98	0.10	-9.69	0.00	-1.18	-0.78
>4	-0.50	0.09	-5.84	0.00	-0.67	-0.33
>5	0.71	0.09	7.61	0.00	0.53	0.89

The estimated threshold (*beta*) values for the items ranged from -2.5 to 1.17, showing significant variability. As an example, the threshold parameters for responses option *strongly agree* across items were (1.13, 1.17, 0.95, and 0.71 for items 1 to 4, respectively) indicating the variability in ISE latent score needed to endorse this response. To illustrate the threshold parameters using the first item, (Table 5), which assesses the struggle with *wanting to do schoolwork*, an average respondent with a latent score of -2.3 had a 50% probability of selecting

disagree or a more affirmative response over *strongly disagree*. At a latent score of -1.10, the respondent was equally likely to choose *disagree* or *neutral*. When the latent score increases to -0.51, *agree* became just as probable as *neutral*. Finally, at a score of 1.13, the probability of an average respondent choosing *strongly agree* was 50%, making it equally likely compared to choosing any less affirmative response.

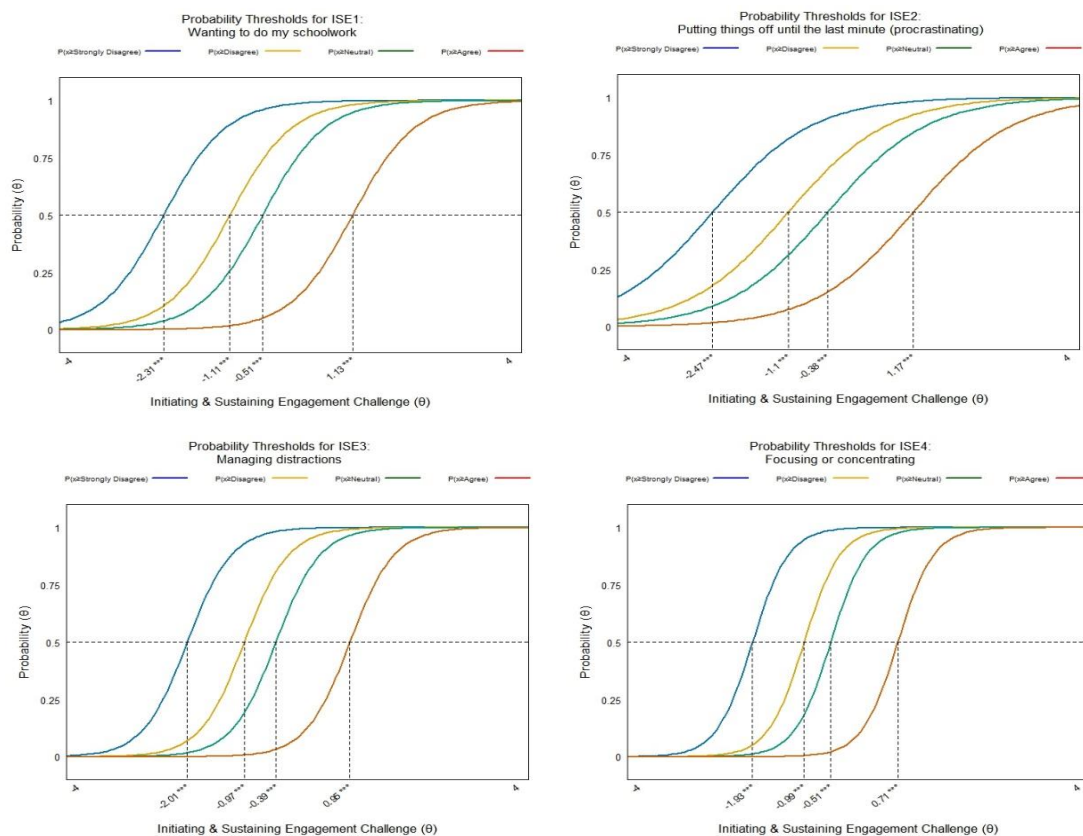
Considering the latent score akin to a z-score which is normally distributed with a mean of zero, a latent score above 1.13 implies a significant deviation above the average for the ISE challenge being measured. So, in a survey context focusing on the past two weeks, someone with a score above 1.13 is likely to select *strongly agree* for an item measuring their struggles with wanting to do schoolwork (ISE1).

Additionally, the differences between threshold parameters for consecutive response categories varied within each item (Figure 13, see also supplementary materials U). For instance, in the first item, the distance between the thresholds for *strongly disagree-disagree* (b_1-b_1) and *disagree-neutral* (b_2-b_3) was 1.20, while the distance between thresholds for *disagree-neutral* (b_2-b_3) and *neutral-agree* (b_3-b_4) was substantially smaller, at 0.60. Consequently, an individual's total score on the ISE challenge scale might differ from scores obtained using Classical Test Theory (CTT), where item scores are assumed to be equally spaced and simply summed (see Willse, 2022).

The Category Characteristic Curves (CCCs) in Figure 14 demonstrated the relationship between participants' locations on the ISE challenge and the probability of choosing specific response options for each item. The points where two adjacent curves intersect reflect the thresholds at which endorsing a higher response option becomes equally probable. For example, for the first item, respondents with a latent score below -2.3 were most likely to respond with *strongly disagree*, while those with scores between -2.3 and -1.10 were most likely to choose *disagree*, and so forth. Additionally, the plots indicated the probability of selecting other options within this score range; for instance, the likelihood of selecting *strongly agree* was close to zero for scores within this range, while the probability of opting for *agree* was non-zero. As shown, item ISE2 had platykurtic probability distributions which indicated lower discriminative power of this item compared to the other items.

Figure 13

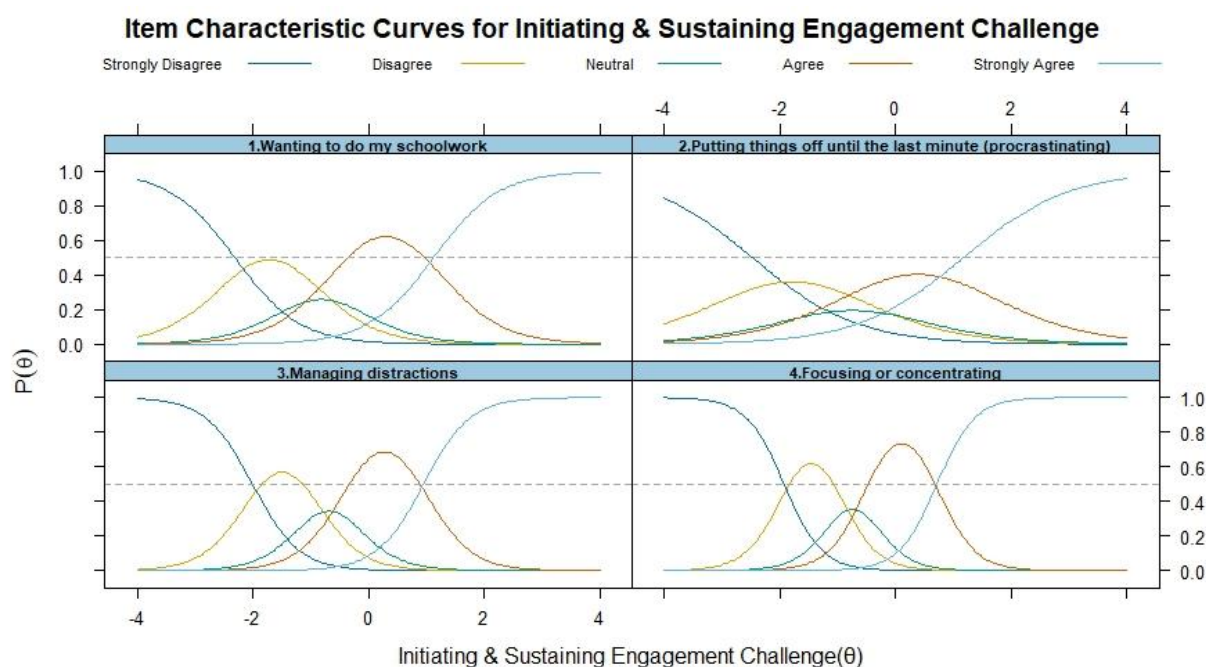
Boundary Characteristic Curves for each Item of the ISE Challenges Scale



Observing the latent score range where participants were inclined to select *neutral* as the midpoint of the Likert scale (a score between -1.10 and -0.51 for ISE1) showed that within this rather restricted range—where the likelihood of choosing *neutral* is at its peak (approximately 25%)—the probabilities of selecting response options *disagree* and *agree* exceeded that of option *neutral*. This observation suggested that the measurement precision at the midpoint of the Likert scale was weaker than the other response categories of the ISE challenge scale. Such a low precision of the midpoint response option was particularly observed for items ISE1 and ISE2.

Figure 14

Category Characteristic Curves for each Item of the ISE Challenges Scale



Item information plots (Figure 15a), integrates threshold and discrimination characteristics to show how much information each item contributes at various latent continuum levels. In other words, these plots helped identify the ranges within the ISE challenge latent continuum that each item accurately measured or failed to cover. Item ISE4 showed the highest level of information, followed by ISE3 and ISE1, while ISE2 had the lowest. Such a low information indicated that this item demonstrated low sensitivity in response to changes of the

latent continuum of ISE challenge. Collectively, the items offered substantial information for measuring the ISE challenge for participants scoring between -2 and 1, with the highest information at a θ value of -0.8. However, information levels decreased as θ values moved beyond this range.

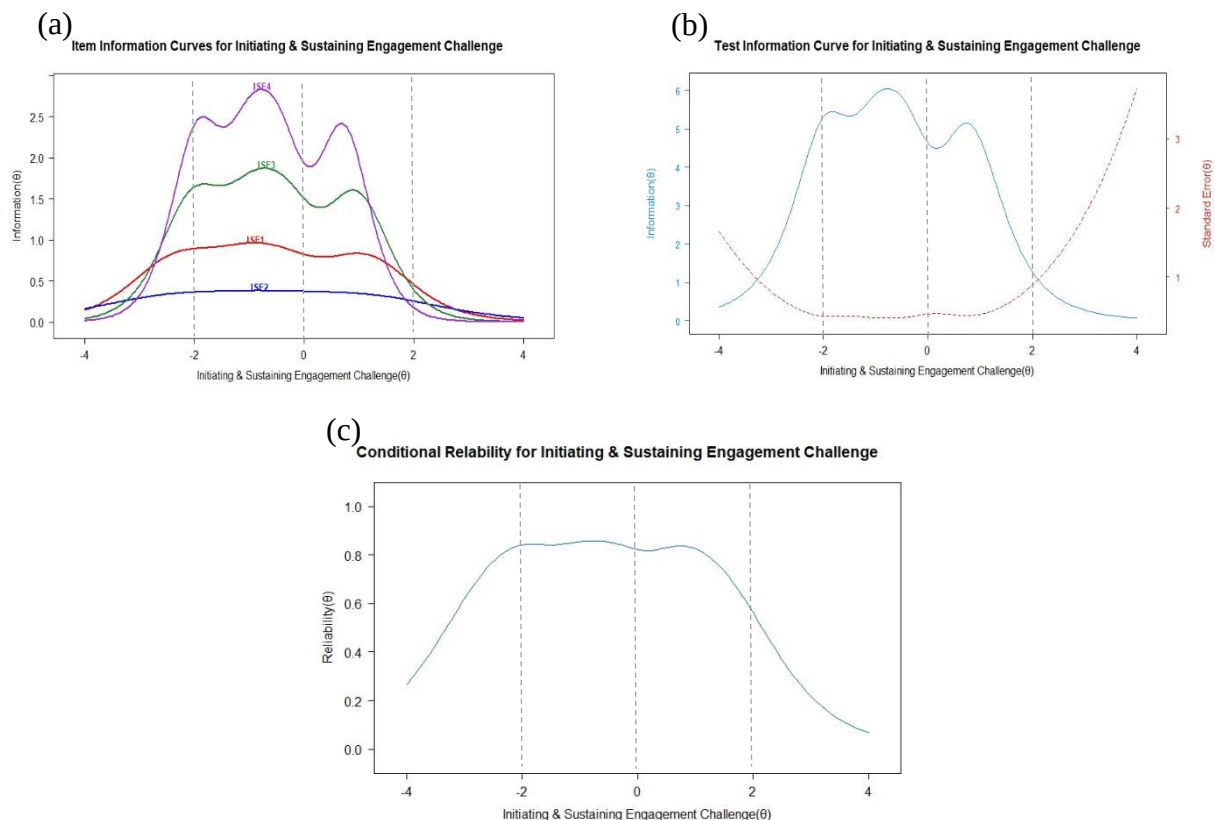
The test information plot (Figure 15b), used to assess the entire scale's measurement precision at various levels of the latent attribute, indicated high reliability and low measurement error within this range. A higher value in the test information plot signifies higher precision in measurement and the standard error of measurement is represented by the red dotted line in the plot. Reliability remained around 0.8 for θ levels from -2 to 1 but drops to 0.6 at θ 2, continuing to decrease as the values deviated further (Figure 15c).

Figure 15

(a) Item Information Function Graphs for the ISE Challenges Items

(b) Information and Standard Error Graphs for the ISE Challenges Scale

(c) Conditional Reliability for the ISE Challenges Scale

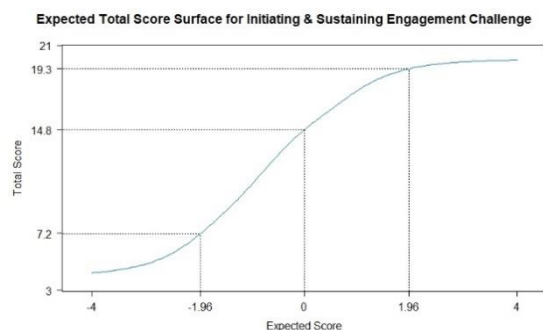


Note. Dotted Vertical lines at $\theta = -2, 0, 2$.

The expected scores from the test characteristic curves plotted against the latent continuum are shown in Figure 16. This plot enables the conversion of theta values into corresponding total scores in the units of the observed variables. By translating latent attribute scores into sum scores, it assists in grasping the distribution and tendencies of responses for the *initiating and sustaining engagement* challenges scale. Given that the ISE challenge scale comprises four items, each with a five-point Likert scale, the range of expected scores falls between 4 and 20. As the plot shows, participants scoring below -1.96 had expected scores of 7.2 or lower, suggesting that they tended to choose the two lowest responses on the Likert scale across items. By applying ± 1.96 , representing two standard deviations in a standard normal distribution, it was estimated that approximately 95% of randomly selected participants would score between 7.2 and 19.3, with an average of 14.8.

Figure 16

Expected Total Score Graph for the ISE Challenges Scale



IRT Analyses of the Goal and Time Management Challenges Scale (GTM)

The comparisons made in the frequency distribution of item responses is helpful because different distributions across items could decrease their linear associations. Frequency distribution for items within goal and time management scale (Table 6) showed that the distribution frequencies varied across the four items. For instance, the frequency of respondents

selecting response option *agree* was 22% for GTM3: struggling with *meeting deadlines*, while it was notably higher, at 39%, for GTM2: struggling with *managing time, tasks, or goals*. The percentage of respondents who chose the lowest response option, *strongly disagree*, varied from 7% for GTM2 to 20% for GTM3. As for the highest response option, *strongly agree*, around 13% of respondents selected this for both GTM1 and GTM3, whereas a larger percentage, 19%, opted for it in the case of GTM4 and GTM2. Notably, the most frequently selected response across categories was option *agree*, with the exception of GTM3, struggling with *meeting deadline*, for which the most selected response was option *disagree*.

Table 6

The Frequency Distributions of the GTM Challenges Items (N=2671)

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
GTM1					
Frequency	224	704	444	951	347
Percent	8.39	26.37	16.63	35.62	13
Cumulative%	8.39	34.76	51.39	87	100
GTM2					
Frequency	182	544	376	1053	515
Percent	6.82	20.37	14.08	39.44	19.29
Cumulative%	6.82	27.19	41.27	80.71	100
GTM3					
Frequency	532	884	315	593	347
Percent	19.92	33.1	11.79	22.2	12.99
Cumulative%	19.92	53.01	64.81	87.01	100
GTM4					
Frequency	216	582	386	978	508
Percent	8.09	21.8	14.46	36.63	19.03
Cumulative%	8.09	29.89	44.34	80.97	100

Regarding item discrimination (*alpha*), all items displayed high and varied parameters, with values ranging from 1.54 to 3.69 (Table 7). According to Baker and Kim (2017), discrimination values between 1.35 to 1.69 can be interpreted as high and above 1.7 as very high. High discrimination values indicate that the items are effective in differentiating between individuals with varying levels of the perceived GTM challenge.

The estimated thresholds (*beta*) varied from -1.7 to 1.66 and were comparatively different in terms of their magnitude. For instance, for the response *strongly agree*, the threshold values were 1.39, 0.95, 1.66, and 1.07 for item1 to 4, respectively. To illustrate the threshold parameters using the first item, struggling with *keeping commitments or goals*, (Table, 7), an average respondent with a latent score of -1.7 had a 50% probability of choosing response 2 (*disagree*) or higher versus choosing any of the lower responses (i.e., *strongly disagree*). An average respondent with a latent score of -0.45 had an equal (50%) chance of selecting response 3 (*neutral*) or higher, as opposed to choosing a lower response category (1 or 2). An average respondent with a latent score of 0.06 had an equal (50%) chance of selecting response 4 (*agree*), or higher versus choosing a lower response category (1, 2, or 3). Finally, an individual with a latent score of 1.39 was equally likely to choose response 5 (*strongly agree*) or a lower response (1, 2, 3, or 4), reflecting a 50% probability in either case.

The latent score can be interpreted as akin to a z-score, which is normally distributed across a population with a mean of zero. If an individual's latent score is above 1.39, they are significantly above the average level for the GTM challenges being measured. In the context of self-reporting experiences over the past two weeks, such a person would be likely to select the highest response option, *strongly agree*, when asked about the extent to which they struggled with *keeping commitments or goals* (GTM1).

Table 7
Item Response Analysis of the GTM Challenges Scale

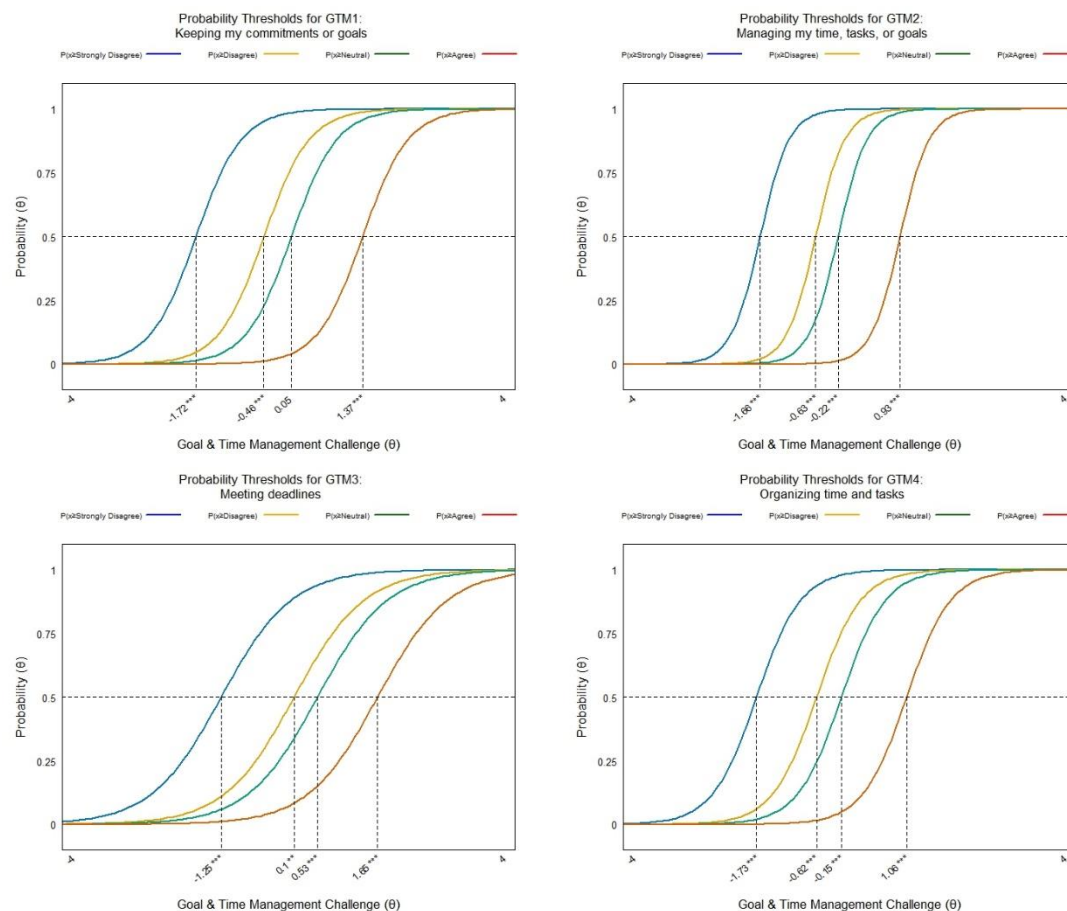
	Coef.	SE	Z	p	CI_2.5	CI_97.5
GTM1: Keeping my commitments or goals						
Discrimination (α)	2.37	0.09	25.95	0.00	2.19	2.55
Thresholds (β)						
>2	-1.71	0.05	-31.19	0.00	-1.82	-1.60
>3	-0.45	0.06	-7.78	0.00	-0.56	-0.33
>4	0.06	0.05	1.34	0.18	-0.03	0.16
>5	1.39	0.37	3.73	0.00	0.66	2.12
GTM2: Managing my time, tasks, or goals						
Discrimination (α)	3.69	0.19	19.43	0.00	3.32	4.07
Thresholds (β)						
>2	-1.66	0.05	-34.22	0.00	-1.75	-1.56
>3	-0.62	0.08	-7.33	0.00	-0.78	-0.45
>4	-0.20	0.07	-2.77	0.01	-0.34	-0.06
>5	0.95	0.36	2.65	0.01	0.25	1.65
GTM3: Meeting deadlines						
Discrimination (α)	1.54	0.06	24.31	0.00	1.42	1.67
Thresholds (β)						
>2	-1.24	0.05	-22.85	0.00	-1.35	-1.13
>3	0.12	0.03	3.57	0.00	0.05	0.18
>4	0.55	0.08	6.81	0.00	0.39	0.71
>5	1.66	0.34	4.94	0.00	1.00	2.32
GTM4: Organizing time and tasks						
Discrimination (α)	2.43	0.09	26.62	0.00	2.25	2.61
Thresholds (β)						
>2	-1.72	0.05	-31.64	0.00	-1.83	-1.62
>3	-0.60	0.06	-9.90	0.00	-0.72	-0.48
>4	-0.14	0.05	-2.74	0.01	-0.24	-0.04
>5	1.07	0.17	6.23	0.00	0.74	1.41

The differences between threshold parameters for consecutive response categories also varied within each item (Figure 17, see also supplementary materials U). For instance, for the first item, the distance between the thresholds for *strongly disagree-disagree* (b1-b2) and *disagree-neutral* (b2-b3) was 1.27, while the distance between thresholds for *disagree-neutral* (b2-b3) and *neutral-agree* (b3-b4) was substantially smaller, at 0.51. Consequently, an

individual's total score on a GTM challenge scale could differ from scores obtained using Classical Test Theory (CTT), where the differences between item responses are assumed to be equal.

Figure 17

Boundary Characteristic Curves for each Item of the GTM Challenges Scale



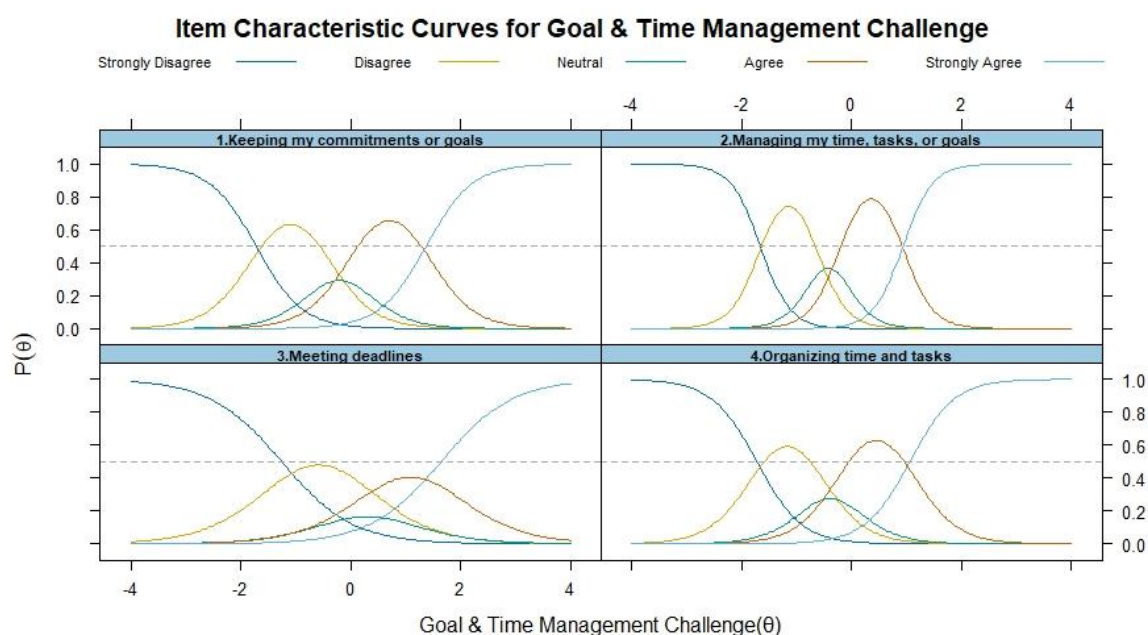
The Category Characteristic Curves (CCCs) in Figure 18 demonstrated the relationship between participants' locations on the GTM challenge and the probability of choosing specific response options for each item. The points at which two adjacent curves intersect indicate the thresholds beyond which selecting a higher response option becomes more probable. For example, for the first item, respondents with a latent score below -1.71 were most likely to

respond with *strongly disagree*, while those with scores between -1.71 and -0.45 were most likely to choose *disagree*, and so forth. Additionally, the curves indicate the probability of selecting other options within this score range; for instance, the likelihood of selecting *strongly agree* was about zero for scores within this range, while the probability of opting for *agree* was non-zero.

Observing the latent score range where participants were inclined to select *neutral* as the midpoint of the Likert scale (a score between -0.45 and -0.06 for GTM1) showed that within this restricted range—where the likelihood of choosing *neutral* is at its peak (approximately 25%)—the probabilities of selecting response options *disagree* and *agree* exceeded that of option *neutral*. This pattern was observed for all items within the scale except GTM2. This observation indicated lower measurement precision for the midpoint of the Likert scale compared to the adjacent response options.

Figure 18

Category Characteristic Curves for each Item of the GTM Challenges Scale



We also examined the item information function for each item (Figure 19a). In IRT, the information function integrates the characteristics of threshold and discrimination parameters to indicate how much information each item contributes at different points along the latent continuum. Essentially, by investigating these plots, we could identify the level on the GTM challenge latent continuum where each item provided (or failed to provide) adequate measurement reliability.

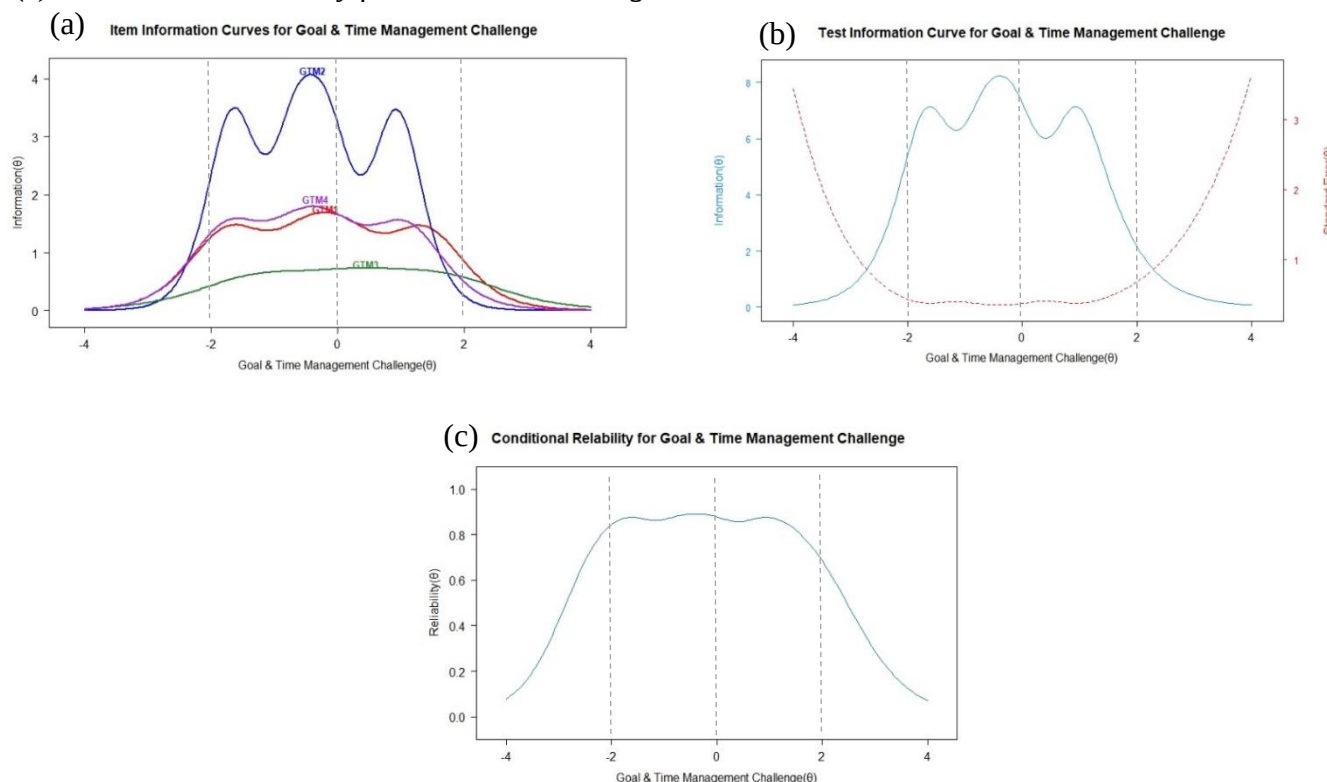
Item GTM2 exhibited a significantly higher level of information, followed by GTM4 and GTM1. In contrast, item GTM3 displayed the lowest level of information and was shown to be less sensitive to changes of the latent continuum of GTM challenges. Overall, all items covered the distribution of the GTM challenge latent continuum (*theta*), spanning a range from -1.8 to 1. Additionally, the plot suggested that reliable information was obtained at an average *theta* value of 0. At this point, item GTM2 yielded an information value of approximately 4, while items GTM1 and GTM3 provided lower values, around 1.5.

The item information functions were aggregated to form the test information function, which was used to assess the precision of the entire scale at various levels of the GTM challenge latent attribute. A higher value in the test information function signifies higher precision in measurement. The scale information plot (Figure 19b) indicated that items collectively offered substantial information for measuring GTM challenge in participants scoring between -1.8 to 1. The standard error of measurement represented by the red dotted line in the plot indicated that the GTM challenge scale demonstrates high reliability and low measurement error within this range. The scale provided the most information and the lowest measurement error slightly below average *theta* levels. However, the information level decreased for *theta* values below -1.8 and above 1.

Figure 19c illustrates the reliability for varying levels of the GTM challenge. As depicted, the reliability remained slightly above 0.8 for θ levels ranging from -2 to 1.8. However, reliability dropped to 0.7 at a θ value of 2 and continued to decrease as the θ values extended beyond this range.

Figure 19

- (a) Item Information Function Graphs for the GTM Challenges Items
 (b) Information and Standard Error Graphs for the GTM Challenges Scale
 (c) Conditional Reliability for the GTM Challenges Scale



Note. Dotted Vertical lines at $\theta = -2, 0$, and 2 .

Figure 20 displays the expected scores derived from the test characteristic curves, plotted against the latent continuum (GTM challenge). Converting latent attribute scores to sum scores helps in understanding the distribution and tendencies of responses using the original metrics used in the GTM challenges scale. Given that the GTM challenges scale comprises four items, each with a five-point Likert scale, the range of expected scores falls between 4 and 20. The

results indicated that participants scoring below -1.96 on the latent continuum would have expected scores of 5.4 or lower, suggesting they are most likely to select response options *strongly disagree* or *disagree* in response to the items. By choosing ± 1.96 , which represents two standard deviations in the standard normal distribution, we could infer that approximately 95% of randomly selected participants would score between 5.4 and 19.1 with an average of 13.

Figure 20

Expected Total Score Graph for the GTM Challenges Scale



IRT Analyses of the Cognitive Challenges Scale (COG)

Frequency distribution tables for all *cognitive challenges* items were examined as the first step (Table 8). Comparing the frequency distributions of responses across items can enhance the understanding of linear associations among them.

The distribution frequencies varied across the six items of the COG challenge. For instance, the frequency of respondents selecting response option *agree* was 39% for COG5, struggling with *remembering things that I learned*, while it was notably lower, at 22%, for COG3, struggling with *language and communication*. The minimum percentage of respondents selecting *strongly disagree* was 6.2% for COG5, while the maximum was 17.5%, for COG3. For response option *strongly agree*, the minimum percentage was approximately 5.7% for COG3, with the maximum being 13.1% for COG5. Notably, across all items, the response option *agree*

was the most frequently selected, except for items COG3 and COG4 for which response option *disagree* was the most frequently selected response option.

Table 8

The Frequency Distributions of the Cognitive Challenges Items (N=2671)

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
COG1					
Frequency	219	751	597	858	239
Percent	8.22	28.19	22.41	32.21	8.97
Cumulative%	8.22	36.41	58.82	91.03	100
COG2					
Frequency	238	762	572	878	216
Percent	8.93	28.58	21.46	32.93	8.1
Cumulative%	8.93	37.51	58.96	91.9	100
COG3					
Frequency	467	965	494	585	153
Percent	17.53	36.22	18.54	21.96	5.74
Cumulative%	17.53	53.75	72.3	94.26	100
COG4					
Frequency	320	943	514	721	172
Percent	11.99	35.32	19.25	27	6.44
Cumulative%	11.99	47.3	66.55	93.56	100
COG5					
Frequency	165	612	498	1046	350
Percent	6.18	22.91	18.64	39.16	13.1
Cumulative%	6.18	29.09	47.73	86.9	100
COG6					
Frequency	217	772	565	890	226
Percent	8.13	28.91	21.16	33.33	8.46
Cumulative%	8.13	37.04	58.2	91.54	100

Regarding item discrimination (*alpha*), all items displayed high to very high slopes, with values ranging from 1.52 to 3.98 (Table 9). According to Baker and Kim (2017), discrimination values between 1.35 to 1.69 can be interpreted as high, and above 1.7 as very high. High discrimination values indicated that the items were effective in differentiating between individuals with varying levels of the perceived COG challenge.

The estimated thresholds (*beta*) values varied from -2.16 to 2.12 and were comparatively different across items in terms of their magnitude (Table 9). For instance, the threshold values for the response option *strongly agree* were 1.44, 1.46, 2.12, 1.81, 1.52, and 1.69 for item1 to 6, respectively. To illustrate the threshold parameters using the first item, struggling with *applying my learning* (Table, 9). An average respondent with a latent score of -1.55 had a 50% probability of choosing response 1 (*strongly disagree*) versus choosing any of the higher responses (2, 3, 4, or 5). An average respondent with a latent score of -0.34 had an equal (50%) chance of selecting either response 1 or 2 (*disagree*), as opposed to choosing a higher response category (3, 4, or 5). An average respondent with a latent score of 0.27 had an equal (50%) chance of selecting response 1, 2, or 3 (*neutral*) versus choosing a higher response category (4 or 5). Finally, an individual with a latent score of 1.44 was equally likely to choose response 5 (*strongly agree*) or any lower response (1, 2, 3, or 4), reflecting a 50% probability in either case. Latent scores (*theta*) can be interpreted akin to z-scores, which are normally distributed across a population with a mean of zero. If an individual's latent score is above 1.44, they are significantly above the average level for the COG challenge latent attribute. Such a person would be likely to select the highest response option, *strongly agree*, when asked about the extent to which they struggled with applying their learning (COG1).

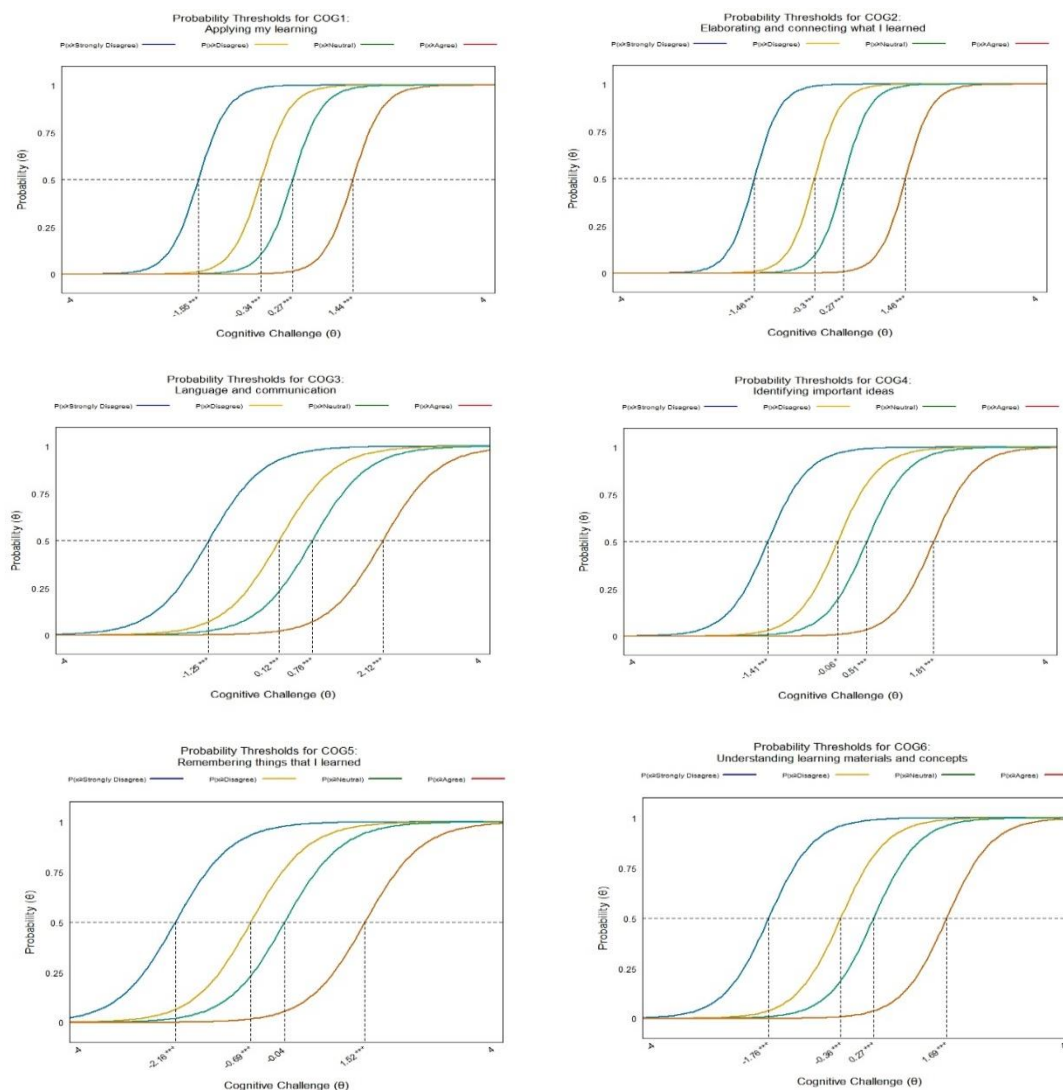
Table 9*Item Response Analysis of the Cognitive Challenges Scale*

	Coef.	SE	Z	p	CI_2.5	CI_97.5
COG1						
Discrimination (α)	3.51	0.13	27.02	0.00	3.25	3.76
Thresholds (β)						
>2	-1.55	0.04	-34.84	0.00	-1.64	-1.47
>3	-0.34	0.03	-12.49	0.00	-0.39	-0.29
>4	0.27	0.03	10.33	0.00	0.22	0.32
>5	1.44	0.04	34.25	0.00	1.36	1.52
COG2						
Discrimination (α)	3.98	0.16	24.86	0.00	3.66	4.29
Thresholds (β)						
>2	-1.46	0.04	-35.18	0.00	-1.55	-1.38
>3	-0.30	0.03	-11.48	0.00	-0.35	-0.25
>4	0.27	0.03	10.34	0.00	0.21	0.32
>5	1.46	0.04	35.06	0.00	1.38	1.54
COG3						
Discrimination (α)	1.88	0.07	27.94	0.00	1.74	2.01
Thresholds (β)						
>2	-1.25	0.05	-26.15	0.00	-1.34	-1.15
>3	0.12	0.03	3.72	0.00	0.06	0.18
>4	0.76	0.04	20.20	0.00	0.68	0.83
>5	2.12	0.07	29.15	0.00	1.98	2.27
COG4						
Discrimination (α)	2.53	0.09	28.86	0.00	2.36	2.70
Thresholds (β)						
>2	-1.41	0.05	-31.08	0.00	-1.50	-1.32
>3	-0.06	0.03	-1.99	0.047	-0.11	0.00
>4	0.51	0.03	16.83	0.00	0.45	0.57
>5	1.81	0.06	32.43	0.00	1.70	1.91
COG5						
Discrimination (α)	1.81	0.06	28.16	0.00	1.69	1.94
Thresholds (β)						
>2	-2.16	0.07	-29.62	0.00	-2.30	-2.01
>3	-0.69	0.04	-18.30	0.00	-0.77	-0.62
>4	-0.04	0.03	-1.15	0.25	-0.10	0.03
>5	1.52	0.05	28.01	0.00	1.42	1.63
COG6						
Discrimination (α)	2.28	0.08	29.17	0.00	2.13	2.44
Thresholds (β)						
>2	-1.76	0.06	-31.86	0.00	-1.87	-1.65
>3	-0.36	0.03	-11.79	0.00	-0.42	-0.30
>4	0.27	0.03	9.22	0.00	0.22	0.33
>5	1.69	0.05	31.35	0.00	1.59	1.80

Additionally, within each item, the differences between threshold parameters for consecutive response categories also varied (Figure 21, see also supplementary materials U). For example, for the first item, the distance between the thresholds for *strongly disagree-disagree* (b1-b2) and *disagree-neutral* (b2-b3) was 1.21, while the distance between thresholds for *disagree-neutral* (b2-b3) and *neutral-agree* (b3-b4) was substantially smaller, at 0.61.

Figure 16

Boundary Characteristic Curves for each Item of the Cognitive Challenges Scale

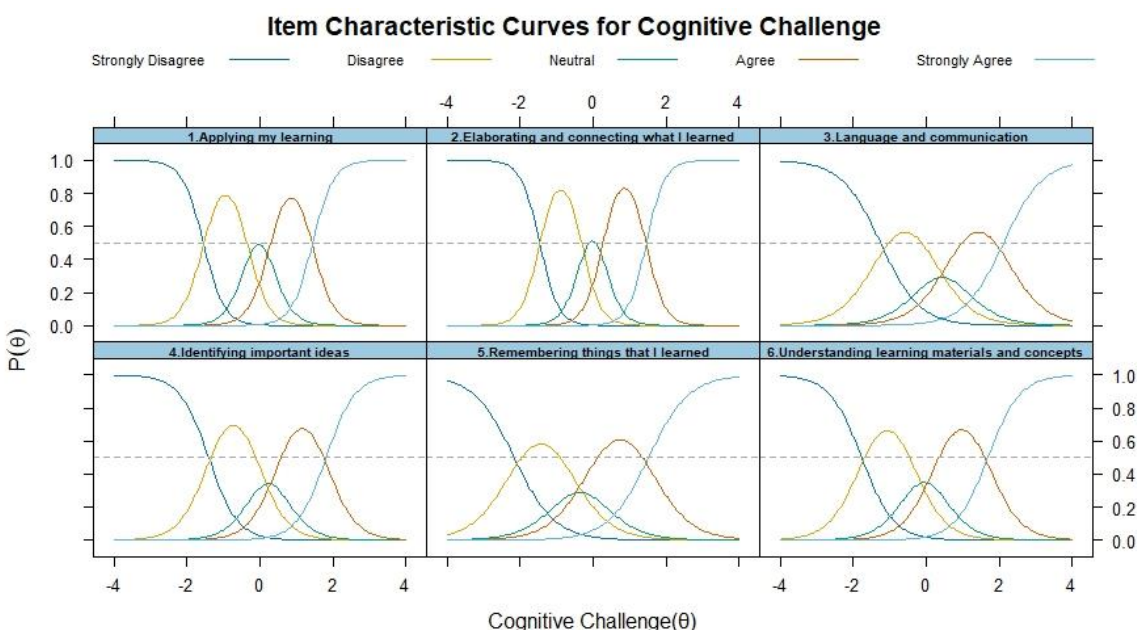


Consequently, an individual's total score on a COG challenge scale could differ from scores obtained using Classical Test Theory (CTT), where such distances between item scores are assumed to be equally spaced (see Willse, 2022).

The Category Characteristic Curves (CCCs) in Figure 21 demonstrated the relationship between participants' locations on the latent cognitive challenge continuum and the probability of choosing specific response options for each item. These curves indicated the thresholds at which a higher response option becomes more probable. The points where two adjacent curves intersect reflect the thresholds at which the transition from one response option to the next occurs. For example, for the first item, respondents with a latent score below -1.55 were most likely to respond with *strongly disagree*, while those with scores between -1.55 and -0.34 were most likely to choose *disagree*, and so forth. Additionally, these curves indicated the probability of selecting other options within this score range; for instance, the likelihood of selecting response option *strongly agree* was close to zero for scores in this range, while the probability of opting for *agree* was non-zero.

Figure 21

Category Characteristic Curves for each Item of the Cognitive Challenges Scale



Observing the latent score range where participants were inclined to select *neutral* as the midpoint of the Likert scale (e.g., a score between -0.34 and 0.27 in item 1) showed that within this restricted range—where the likelihood of choosing *neutral* is at its peak (approximately 45%)—the probabilities of selecting response options *disagree* and *agree* did not exceed that of *neutral*. This characteristic was not observed for items COG 3 and COG5. This observation suggested that the measurement precision at the midpoint of the Likert scale was superior compared to other items evaluated so far, but it remained less than ideal particularly for items COG3 and COG5.

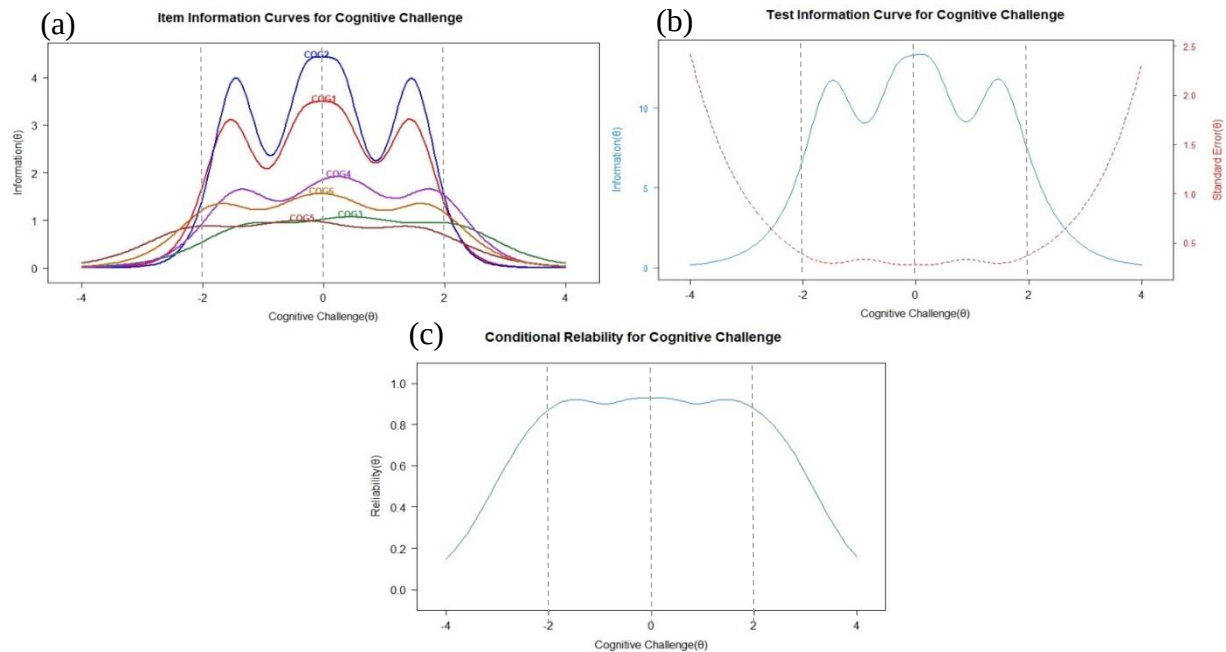
We also examined the item information function for each item (see Figure 22a). In IRT, the information function integrates the item parameters to indicate how much information each item contributes at different points along the latent continuum. Essentially, by investigating these plots, we could identify the level on the *cognitive challenge* continuum where each item provided precise measurement. The standard error of measurement is represented by the red dotted line in this plot. Item COG2 exhibited a significantly higher level of information, followed by COG1, COG4, and COG6. In contrast, items COG3 and COG5 displayed the lowest level of information and were shown to be less sensitive to changes of the latent continuum of COG challenge. Overall, items collectively covered the distribution of the COG challenges latent continuum (*theta*), spanning a range from -1.5 to 1.5. Additionally, the plot suggested that reliable information was obtained at an average *theta* value of 0. At this point, item COG2 yielded an information value of above 4, while items COG1, COG4, and COG6 provide lower values, around 3.2, 1.5, and 1.3, respectively. The information values for items COG 3 and COG5 was approximately 1.

The item information functions were aggregated to form the test information function, which was used to assess the precision of the entire COG challenges scale at various levels of the latent attribute. A higher value in the test information function signifies higher precision in

measurement and the standard error of measurement is represented by the red dotted line in the plot. As shown in Figure 22b, the six items collectively offered substantial information for

Figure 22

- (a) Item Information Function Graphs for the Cognitive Challenges Items
 (b) Information and Standard Error Graphs for the Cognitive Challenges Scale
 (c) Conditional Reliability for the Cognitive Challenges Scale



Note. Dotted Vertical lines at $\theta = -2, 0$, and 2 .

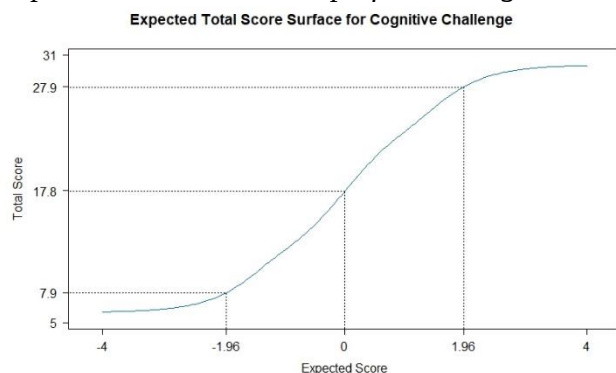
measuring COG challenge in participants scoring between -1.8 and 1.8 . This indicated that the COG challenge scale demonstrated relatively high reliability and low measurement error within this range. The scale provided the most information and the lowest measurement error around the average θ level. However, the information level decreased for θ values below -1.8 and above 1.8 . Figure 22c illustrates the reliability of the scale across varying levels of the COG challenge.

As depicted, the reliability remained around 0.9 for θ levels ranging from -1.8 to 1.8 and still above $.8$ for θ levels ranging from -2 to 2 . However, reliability dropped as the θ values moved away from this range. Overall, the plot indicated high reliability and low measurement error of the COG challenge scale within this range.

Figure 23 displays the expected scores derived from the test characteristic curves, plotted against the latent continuum (COG challenge). Translating latent attribute scores into sum scores helps in understanding the distribution and tendencies of responses using the original metrics of the scale. Given that the COG challenge scale comprises six items, each measured by a five-point Likert scale, the range of expected scores falls between 6 and 30. The plot indicated that participants scoring below -1.96 on the latent continuum would have expected scores of 7.9 or lower, suggesting they were most likely to select response options 1 or 2 on the Likert scale. By choosing ± 1.96 , which represents two standard deviations in the standard normal distribution, we could infer that approximately 95% of randomly selected participants would score between 7.9 and 27.9 with an average of 17.8.

Figure 23

Expected Total Score Graph for the Cognitive Challenges Scale



IRT Analyses of the Metacognitive Challenges Scale (MET)

Frequency distribution tables for all *metacognitive challenges* items were examined as the first step (Table 10). The comparisons made in the frequency distribution of item responses is helpful because one of the reasons that can reduce the linear associations among items is their varying distributions. The distribution frequencies varied across the eight items of the MET challenge scale. For instance, the frequency of respondents selecting response option 4 (*agree*) was 33.8% for MET4, struggling with *knowing how or when to fix strategies/study skills*, while it was lower, at 28%, for MET2, struggling with *interpreting tasks and expectations*.

The minimum percentage of respondents selecting the response option *strongly disagree* was 8.3% for MET4, while the maximum was 10.2% for MET5. For the *strongly agree* option, the minimum/maximum range was 7.8% for MET2 to 12% for MET7. Notably, across all items, the response option *agree* was the most frequently selected option except for item MET2 for which the response option *disagree* was the most frequently selected option. This observation indicated that across all items, struggling with *interpreting tasks and expectations* was perceived as the least challenging item. To elaborate, 64% of respondents rated this item, as *neutral* or lower, while the corresponding frequency was between 53.9% to 59.6% for the other items.

Table 10
The Frequency Distributions of the Metacognitive Challenges Items (N=2671)

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
MET1					
Frequency	237	774	580	843	235
Percent	8.88	29	21.73	31.58	8.8
Cumulative%	8.88	37.88	59.61	91.2	100
MET2					
Frequency	241	850	619	748	209
Percent	9.04	31.87	23.21	28.05	7.84
Cumulative%	9.04	40.91	64.12	92.16	100
MET3					
Frequency	230	818	527	873	220
Percent	8.62	30.66	19.75	32.72	8.25
Cumulative%	8.62	39.28	59.03	91.75	100
MET4					
Frequency	220	728	569	902	246
Percent	8.26	27.32	21.35	33.85	9.23
Cumulative%	8.26	35.57	56.92	90.77	100
MET5					
Frequency	273	841	431	855	269
Percent	10.23	31.51	16.15	32.03	10.08
Cumulative%	10.23	41.74	57.89	89.92	100
MET6					
Frequency	257	783	547	851	230
Percent	9.63	29.35	20.5	31.9	8.62
Cumulative%	9.63	38.98	59.48	91.38	100
MET7					
Frequency	268	740	430	911	320
Percent	10.04	27.73	16.11	34.13	11.99
Cumulative%	10.04	37.77	53.88	88.01	100
MET8					
Frequency	229	718	591	874	255
Percent	8.59	26.92	22.16	32.77	9.56
Cumulative%	8.59	35.51	57.67	90.44	100

Regarding item discrimination (*alpha*), all items displayed very high slopes, with values ranging from 2.44 to 3.48 (Table 11). According to Baker and Kim (2017), discrimination values above 1.7 can be interpreted as very high. High discrimination values indicated that the items were effective in differentiating between individuals with varying levels of the perceived MET challenge.

The estimated thresholds (*beta*) varied from -1.49 to 1.45. For instance, for the response *strongly agree*, the threshold values were 1.44, 1.45, 1.40, 1.34, 1.41, 1.37, 1.33, and 1.37 for items 1 to 8, respectively. To illustrate the threshold parameters using the first item, struggling with *breaking tasks down* (Table 11).

An average respondent with a latent score of -1.46 had a 50% probability of choosing response 2 (*disagree*) or higher versus choosing *strongly disagree*. An average respondent with a latent score of -0.28 had an equal (50%) chance of selecting response option 3 (*neutral*) or higher, as opposed to choosing a lower response category (1 or 2). An average respondent with a latent score of 0.27 had an equal (50%) chance of selecting response 4 (*agree*) versus choosing a lower response category (1, 2, or 3). Finally, an individual with a latent score of 1.44 was equally likely to choose response 5 (*strongly agree*) as they were to choose any lower response (1, 2, 3, or 4), reflecting a 50% probability in either case. Given that a latent score can be interpreted as akin to a z-score, which is normally distributed across a population with a mean of zero, if an individual's latent score is above 1.44, they are significantly above the average level for the MET challenge. In the context of survey questions about their experiences over the past two weeks, such a person would be likely to select the highest response option, *strongly agree*, when asked about the extent to which they struggled with breaking tasks down (MET1).

Table 11

Item Response Analysis of the Metacognitive Challenges Scale

	Coef.	SE	Z	p	CI_2.5	CI_97.5
MET1						
Discrimination (α)	2.96	0.09	32.50	0.00	2.78	3.14
Thresholds (β)						
>2	-1.46	0.04	-34.85	0.00	-1.54	-1.38
>3	-0.28	0.04	-6.89	0.00	-0.36	-0.20
>4	0.27	0.04	6.70	0.00	0.19	0.35
>5	1.44	0.80	1.80	0.07	-0.12	3.00
MET2						
Discrimination (α)	3.26	0.10	32.00	0.00	3.06	3.46
Thresholds (β)						
>2	-1.41	0.04	-35.54	0.00	-1.49	-1.34
>3	-0.21	0.04	-5.16	0.00	-0.28	-0.13
>4	0.37	0.04	8.38	0.00	0.28	0.45
>5	1.45	1.26	1.15	0.25	-1.02	3.92
MET3						
Discrimination (α)	3.48	0.11	31.56	0.00	3.26	3.69
Thresholds (β)						
>2	-1.42	0.04	-35.89	0.00	-1.50	-1.35
>3	-0.24	0.04	-5.67	0.00	-0.32	-0.15
>4	0.24	0.04	5.92	0.00	0.16	0.32
>5	1.40	1.29	1.08	0.28	-1.14	3.94
MET4						
Discrimination (α)	3.40	0.11	32.15	0.00	3.19	3.61
Thresholds (β)						
>2	-1.47	0.04	-35.80	0.00	-1.55	-1.39
>3	-0.33	0.04	-7.41	0.00	-0.42	-0.24
>4	0.19	0.04	4.87	0.00	0.12	0.27
>5	1.34	0.93	1.44	0.15	-0.48	3.16
MET5						
Discrimination (α)	2.61	0.08	31.74	0.00	2.45	2.77
Thresholds (β)						
>2	-1.42	0.04	-33.35	0.00	-1.50	-1.34
>3	-0.20	0.04	-5.37	0.00	-0.27	-0.12
>4	0.23	0.04	5.83	0.00	0.15	0.31
>5	1.41	0.49	2.86	0.00	0.44	2.38
MET6						
Discrimination (α)	3.54	0.11	31.13	0.00	3.32	3.77
Thresholds (β)						
>2	-1.34	0.04	-35.59	0.00	-1.41	-1.27
>3	-0.25	0.04	-6.25	0.00	-0.32	-0.17
>4	0.25	0.04	6.54	0.00	0.17	0.32
>5	1.37	1.26	1.09	0.28	-1.10	3.84
MET7						
Discrimination (α)	2.44	0.08	30.98	0.00	2.29	2.60
Thresholds (β)						
>2	-1.46	0.05	-32.45	0.00	-1.55	-1.37
>3	-0.31	0.04	-7.91	0.00	-0.39	-0.24
>4	0.12	0.04	3.22	0.00	0.05	0.20
>5	1.33	0.32	4.09	0.00	0.69	1.96
MET8						
Discrimination (α)	3.03	0.09	32.11	0.00	2.84	3.21
Thresholds (β)						
>2	-1.49	0.04	-35.04	0.00	-1.57	-1.40
>3	-0.34	0.04	-7.91	0.00	-0.43	-0.26
>4	0.22	0.04	5.56	0.00	0.14	0.30
>5	1.37	0.68	2.00	0.045	0.03	2.71

The difference between threshold parameters for consecutive response categories varied within each item (Figure 24, see also supplementary materials U). For instance, in the first item, the distance between the thresholds for *strongly disagree-disagree* (b1-b2) and *disagree-neutral* (b2-b3) was 1.18, while the distance between thresholds for *disagree-neutral* (b2-b3) and *neutral-agree* (b3-b4) was substantially smaller, at -0.28. Consequently, an individual's total score on the MET challenge scale might differ from scores obtained using Classical Test Theory (CTT), where item scores are simply summed as they are assumed to be equally spaced (see Willse, 2022).

The Category Characteristic Curves (CCCs) in Figure 25 demonstrated the relationship between participants' locations on the MET challenge and the probability of choosing specific response options for each item. The points where two adjacent curves intersect reflect the thresholds at which endorsing a higher response option becomes equally probable. For example, for the first item, respondents with a latent score below -1.46 were most likely to respond with *strongly disagree*, while those with scores between -1.46 and -0.28 were most likely to choose *disagree*, and so forth. Additionally, the plots indicated the probability of selecting other options within this score range; for instance, the likelihood of selecting *strongly agree* was close to zero for scores within this range, while the probability of opting for *agree* was non-zero. All MET challenge items exhibited noticeably steep slopes.

Observing the latent score range where participants were inclined to select *neutral* as the midpoint of the Likert scale (a score between -0.28 and 0.27 for MET1) showed that within this restricted range—where the likelihood of choosing *neutral* is at its peak (approximately 40%)—the probabilities of selecting response options *disagree* and *agree* did not exceed that of option *neutral*. This observation suggested that the measurement precision at the midpoint of the Likert scale was better than that of the other scales but still lower than the other response categories of the MET challenge scale. Such a low precision of the midpoint response option was particularly observed for items MET5 and MET7.

Figure 24

Boundary Characteristic Curves for each Item of the Metacognitive Challenges Scale

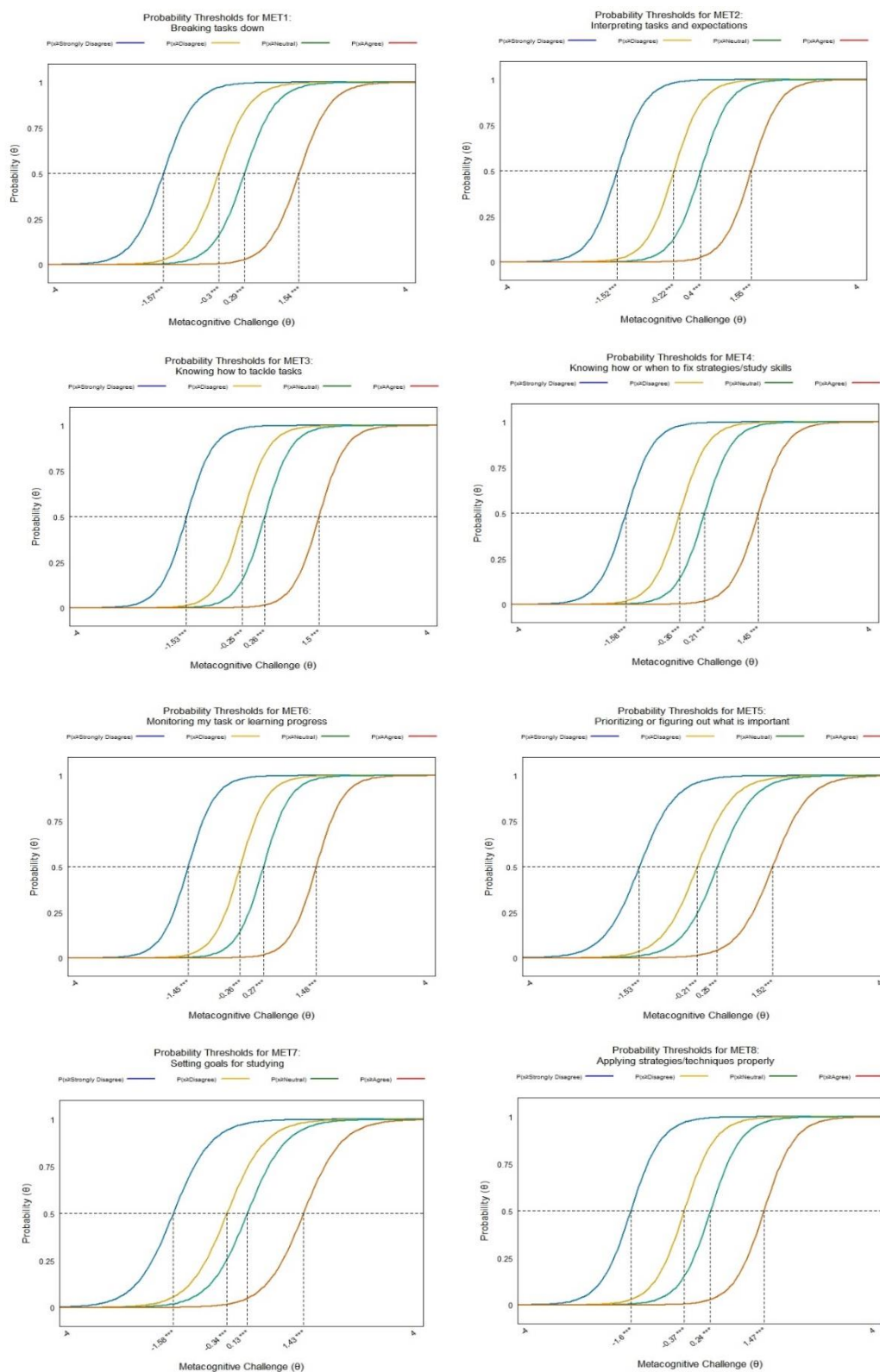
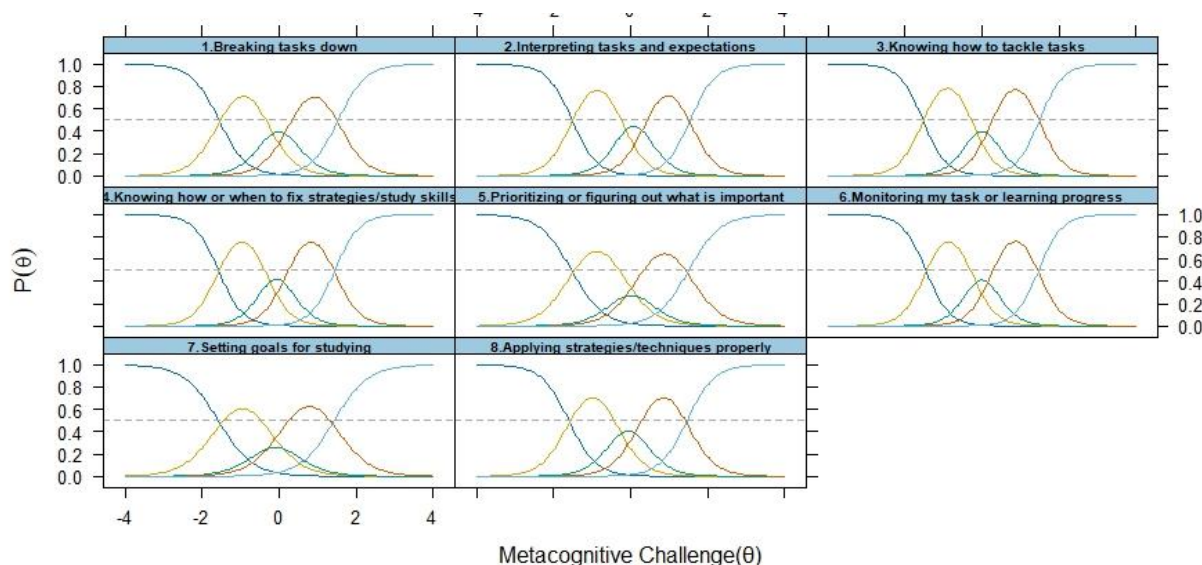


Figure 25
Category Characteristic Curves for each Item of the Metacognitive Challenges Scale



We also examined the item information function for each item (see Figure 26a). The information function integrated the characteristics of threshold and discrimination parameters to indicate how much information each item contributes at different points along the MET challenge latent continuum. Essentially, by investigating this plot, we identified the levels on the MET challenge latent continuum which were reliably measured or left unmeasured by each item.

Item MET3 and MET6 exhibited the highest level of information, followed by MET4, MET2, MET8, MET1. In contrast, items MET5 and MET7 displayed the lowest level of information and exhibited lower sensitivity in response to changes of the latent continuum of *metacognitive* challenge compared to the other 6 items. Overall, all MET items, covered the distribution of the *metacognitive* challenge latent continuum (*theta*), spanning a range from -1.5 to 1.5. Additionally, the plot suggested that reliable information was obtained at an average *theta* value of 0 ranging from 3.2 for MET3 and MET6 to 1.5 for MET7.

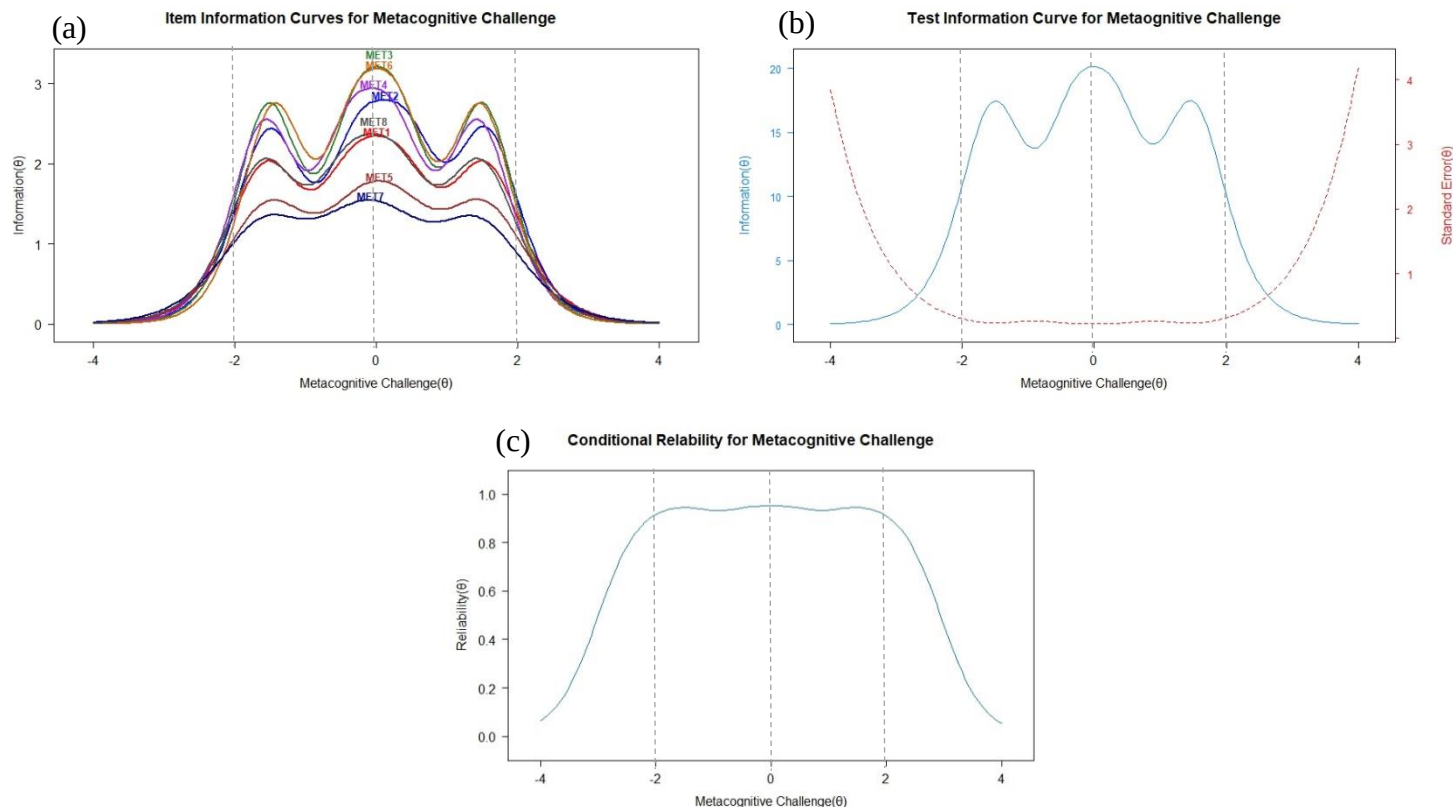
The item information functions were aggregated to form the test information function, which was used to assess the precision of the entire MET challenge scale at various levels of the latent attribute. A higher value in the test information function signifies higher precision in measurement and the standard error of measurement is represented by the red dotted line in the

plot. As shown in Figure 26b, the eight items collectively offered substantial information for measuring MET challenge in participants scoring between -1.5 and 1.5. This indicated that the MET challenge scale demonstrated relatively high reliability and low measurement error within this range. The scale provided the most information and the lowest measurement error around the average theta level. However, the information level decreased for θ values below -1.5 and above 1.5.

Figure 26c illustrates the reliability for varying levels of the MET challenge. As depicted, the reliability remains around 0.95 for θ levels ranging from -1.8 to 1.8 and around .9 for θ levels ranging from -2 to 2. However, reliability dropped as the θ values moved away from this range.

Figure 26

- (a) Item Information Function Graphs for the Metacognitive Challenges Items
 (b) Information and Standard Error Graphs for the Metacognitive Challenges Scale
 (c) Conditional Reliability for the Metacognitive Challenges Scale

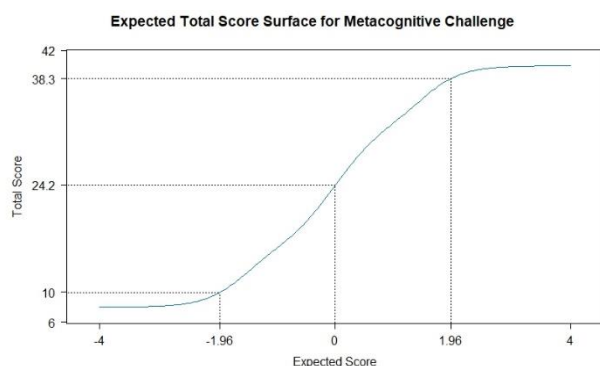


Note. Dotted Vertical lines at $\theta = -2, 0, \text{ and } 2$.

Figure 27 displays the expected scores derived from the test characteristic curves, plotted against the latent continuum (MET challenge). This plot can be used to convert *theta* values into corresponding total scores in the unit of the observed variables. Converting latent attribute scores to sum scores helps in understanding the distribution and tendencies of responses within the MET challenge context. Given that the MET challenge scale comprises eight items, each with a five-point Likert scale, the range of expected scores falls between 8 and 40. The results indicated that participants scoring below -1.96 on the latent continuum would have expected scores of 10 or lower, suggesting they are most likely to select response options 1 or 2 on the Likert scale. By choosing ± 1.96 , which represents two standard deviations in the standard normal distribution, we could infer that approximately 95% of randomly selected participants would score between 10 and 38.3 with an average of 24.2.

Figure 27

Expected Total Score Graph for the Metacognitive Challenges Scale



IRT Analyses of the Social and Emotional Challenges Scale (SE)

Frequency distribution tables for all *social and emotional challenges* items were examined as the first step (Table 12). Comparing the frequency distributions of item responses aids in deepening our understanding of the linear associations among items. The distribution frequencies varied across the six items of the SE challenge scale. For instance, the frequency of respondents selecting response option *agree* was 36% for SE1, struggling with *feeling connected* while it was lower, at 29%, for SE5 struggling with *managing relationships*.

Table 12*The Frequency Distributions of the SE Challenges Items (N=2671)*

	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
SE1					
Frequency	201	621	499	971	369
Percent	7.55	23.34	18.75	36.49	13.87
Cumulative%	7.55	30.89	49.64	86.13	100
SE2					
Frequency	258	598	454	821	531
Percent	9.69	22.46	17.05	30.84	19.95
Cumulative%	9.69	32.16	49.21	80.05	100
SE3					
Frequency	250	633	545	809	428
Percent	9.38	23.75	20.45	30.36	16.06
Cumulative%	9.38	33.13	53.58	83.94	100
SE4					
Frequency	212	596	500	921	437
Percent	7.95	22.36	18.75	34.55	16.39
Cumulative%	7.95	30.31	49.06	83.61	100
SE5					
Frequency	307	727	546	780	301
Percent	11.54	27.32	20.52	29.31	11.31
Cumulative%	11.54	38.86	59.38	88.69	100
SE6					
Frequency	220	535	468	914	533
Percent	8.24	20.04	17.53	34.23	19.96
Cumulative%	8.24	28.28	45.81	80.04	100

The minimum percentage of respondents selecting *strongly disagree* was 7.6% for SE1, while the maximum was 11.5% for SE5. For the response option *strongly agree*, the range was 11.3% for SE5 to 20% for SE2 and SE6. Notably, across all items, the response option *agree* was the most frequently selected response option. 59.4% of respondents rated SE5 (struggling with *managing relationships*) as *neutral* or lower, while 45.8% of them rated SE6 (struggling with *taking care of my mental health and well-being*) as *neutral* or lower. This indicated that, on average, students viewed SE6 as being more challenging than SE5.

Regarding item discrimination (*alpha*), all items displayed moderate to very high slopes, with values ranging from 1.52 to 3.98 (Table 13). According to Baker and Kim (2017), discrimination values ranging from 0.65 to 1.34 can be interpreted as moderate, values from 1.35 to 1.69 as high, and values above 1.7 as very high. High discrimination values indicated that the items were effective in differentiating between individuals with varying levels of the perceived SE challenge. Across the items, item SE2 exhibited the least discrimination.

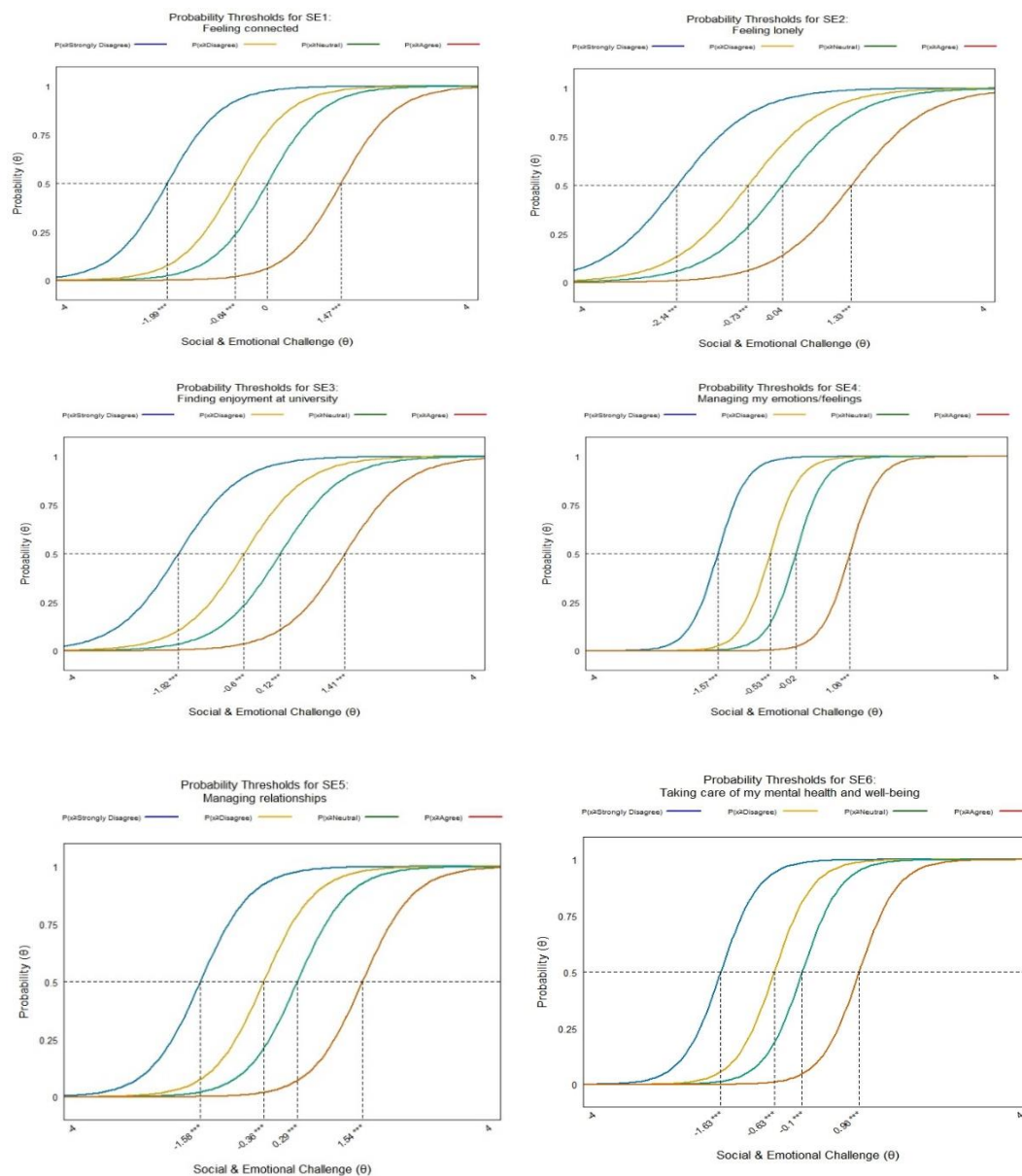
The estimated thresholds (*beta*) varied from -2.13 to 1.58 and were comparatively different across items. For example, for the response *strongly agree* the threshold values were 1.5, 1.37, 1.44, 1.09, 1.58, and 1 for items SE1 to SE6, respectively. To illustrate the threshold parameters using the first item, struggling with *feeling connected* (Table, 13). An average respondent with a latent score of -1.98 had a 50% probability of choosing response 1 (*strongly disagree*) versus choosing any of the higher responses (2, 3, 4, or 5). An average respondent with a latent score of -0.62 had an equal (50%) chance of selecting either response 1 or 2 (*disagree*), as opposed to choosing a higher response category (3, 4, or 5). An average respondent with a latent score of 0.03 had an equal (50%) chance of selecting response 1, 2, or 3 (*strongly disagree* to *neutral*) versus choosing a higher response category (4 or 5). Finally, an individual with a latent score of 1.50 was equally likely to choose response 5 (*strongly agree*) as they were to choose any lower response (1, 2, 3, or 4), reflecting a 50% probability in either case. Given the latent scores can be interpreted as akin to a z-score, which is normally distributed across a population with a mean of zero, if an individual's latent score is above 1.50, they are significantly above the average level for the SE challenge. In the context of survey questions about their experiences over the past two weeks, such a person would be likely to select the highest response option, *strongly agree*, when asked about the extent to which they struggled with feeling connected (SE1).

Within each item, the differences between threshold parameters for consecutive response categories also varied (Figure 28, also see supplementary materials U). For instance, for the first item, the distance between the thresholds for *strongly disagree-disagree* (b_1-b_2) and *disagree-neutral* (b_2-b_3) was 1.36, while the distance between thresholds for *disagree-neutral* (b_2-b_3) and *neutral-agree* (b_3-b_4) was substantially smaller, at 0.65. Consequently, an individual's total score on a SE challenge scale could differ from scores obtained using Classical Test Theory (CTT), where it is assumed that the differences between item scores are equally spaced and can be summed through simple aggregation.

Table 13*Item Response Analysis of the SE Challenges Scale*

	Coef.	SE	Z	p	CI_2.5	CI_97.5
SE1						
Discrimination (α)	1.83	0.07	26.91	0.00	1.69	1.96
Thresholds (β)						
>2	-1.98	0.07	-27.87	0.00	-2.12	-1.84
>3	-0.62	0.06	-9.78	0.00	-0.74	-0.50
>4	0.03	0.05	0.53	0.59	-0.08	0.13
>5	1.50	0.26	5.81	0.00	1.00	2.01
SE2						
Discrimination (α)	1.31	0.05	24.15	0.00	1.20	1.41
Thresholds (β)						
>2	-2.13	0.09	-23.87	0.00	-2.31	-1.96
>3	-0.70	0.07	-10.55	0.00	-0.83	-0.57
>4	-0.01	0.06	-0.24	0.81	-0.12	0.10
>5	1.37	0.15	9.05	0.00	1.07	1.66
SE3						
Discrimination (α)	1.61	0.06	26.26	0.00	1.49	1.73
Thresholds (β)						
>2	-1.90	0.07	-26.00	0.00	-2.05	-1.76
>3	-0.58	0.06	-9.83	0.00	-0.69	-0.46
>4	0.15	0.05	2.96	0.00	0.05	0.25
>5	1.44	0.21	6.91	0.00	1.03	1.85
SE4						
Discrimination (α)	3.45	0.14	24.52	0.00	3.17	3.72
Thresholds (β)						
>2	-1.56	0.05	-31.02	0.00	-1.66	-1.46
>3	-0.51	0.06	-8.04	0.00	-0.63	-0.38
>4	0.01	0.05	0.23	0.81	-0.09	0.11
>5	1.09	0.46	2.39	0.02	0.20	1.99
SE5						
Discrimination (α)	2.03	0.07	27.70	0.00	1.89	2.17
Thresholds (β)						
>2	-1.56	0.06	-27.35	0.00	-1.67	-1.45
>3	-0.33	0.05	-7.10	0.00	-0.42	-0.24
>4	0.32	0.05	6.82	0.00	0.23	0.41
>5	1.58	0.41	3.80	0.00	0.76	2.39
SE6						
Discrimination (α)	2.76	0.10	27.23	0.00	2.56	2.96
Thresholds (β)						
>2	-1.62	0.05	-30.02	0.00	-1.73	-1.52
>3	-0.60	0.06	-10.24	0.00	-0.72	-0.49
>4	-0.07	0.05	-1.54	0.12	-0.17	0.02
>5	1.00	0.18	5.51	0.00	0.64	1.35

Figure 28
Boundary Characteristic Curves for each Item of the SE Challenges Scale

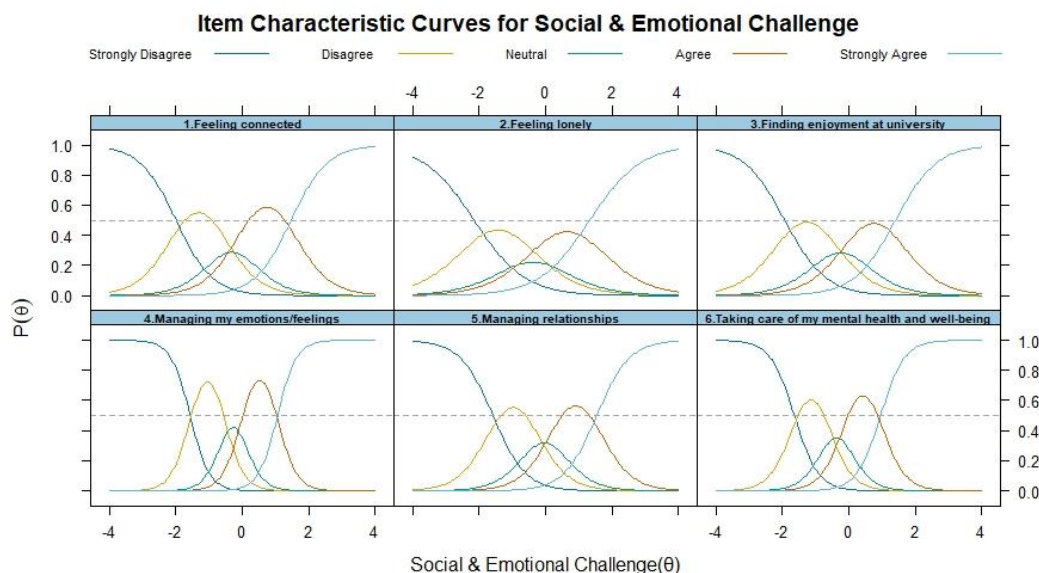


The Category Characteristic Curves (CCCs) in Figure 29 demonstrated the relationship between participants' locations on the SE challenge and the probability of choosing specific response options for each item. These curves indicated the thresholds beyond which a higher response option becomes more probable. These thresholds were characterized as the points where two adjacent curves intersect. For example, for the first item, respondents with a latent

score below -1.98 were more likely to respond with *strongly disagree*, while those with scores between -1.98 and -0.62 were most likely to choose *disagree*, and so forth. Additionally, the curves indicated the probability of selecting other options within this score range; for instance, the likelihood of selecting *strongly agree* is close to zero in this range, while the probability of opting for *agree* is non-zero. Observing the latent score range where participants are inclined to select *neutral* as the midpoint of the Likert scale (a score between -0.62 and 0.03) showed that within this restricted range—where the likelihood of choosing *neutral* was at its peak (approximately 30%)—the probabilities of selecting response options *disagree* and *agree* were also relatively high, exceeding that of option *neutral*. This suggested lower measurement precision at the midpoint of the Likert scale compared to other response options, particularly for items SE1, SE2, and SE3.

Figure 29

Boundary Characteristic Curves for each Item of the SE Challenges Scale



We also examined the item information function for each item (see Figure 30a). In IRT, the information function integrates item characteristics to represent how much information each item contributes at different points along the latent continuum. Essentially, by investigating these plots, we could identify the level on the SE challenge latent continuum where each item provided reliable measurement or failed to measure reliably.

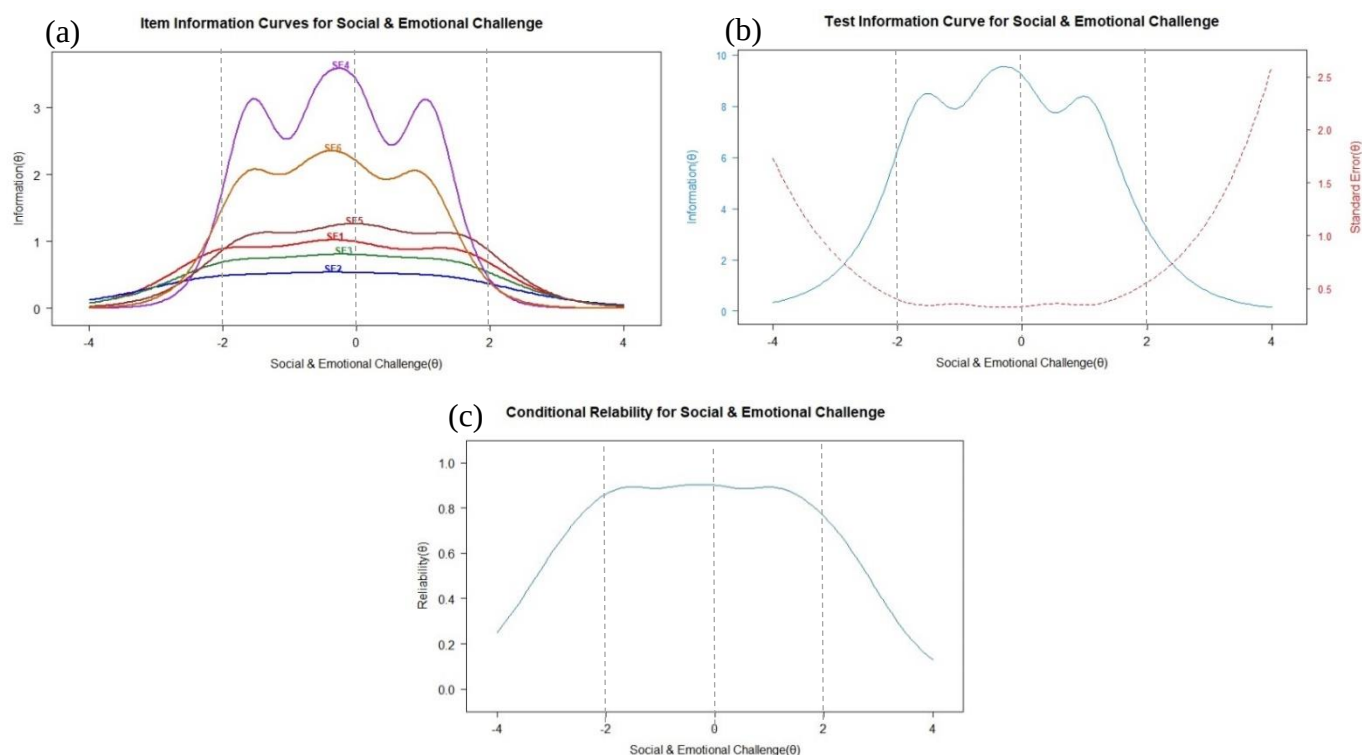
Item SE4 exhibited a significantly higher level of information, followed by SE6 and SE5. In contrast, compared to other items, SE1, SE3, and SE2 displayed lower level of information and were less sensitive to changes of the latent continuum of SE challenge. Overall, items SE4, SE6, and SE5, covered the distribution of the SE challenge latent continuum (θ), spanning a range from -1.6 to 1.1. Additionally, the plot suggested that reliable information was obtained at an average θ value of 0. At this point, item SE4 yielded an information value of above 3, while items SE6 and SE5, and SE1 provided lower values, around 2.4, 1.4, and 1 respectively. The information values for SE3 and SE2 were below 1.

The item information functions were aggregated to form the test information function, which was used to assess the precision of the entire scale at various levels of the SE challenge latent continuum. A higher value in the test information function signifies higher precision in measurement and the standard error of measurement is represented by the red dotted line in the plot. Figure 30b, indicated that the six items collectively offered substantial information for measuring SE challenge in participants scoring between -1.8 and 1.1. Overall, the plot indicated that the SE challenge scale demonstrated low measurement error within this range. The scale provided the most information and the lowest measurement error around the θ level at -0.2. However, the information level decreased for θ values below -1.8 and above 1.1.

Figure 30c illustrates the reliability for varying levels of the SE challenge. As depicted, the reliability remains around 0.9 for theta levels ranging from -1.9 to 1.3. The minimum reliability for *theta* levels ranging from -2 to 2 was about .8. However, reliability dropped for *theta* values beyond this range.

Figure 30

- (a) Item Information Function Graphs for the SE Challenges Items
 (b) Information and Standard Error Graphs for the SE Challenges Scale
 (c) Conditional Reliability for the Motivation SE Scale



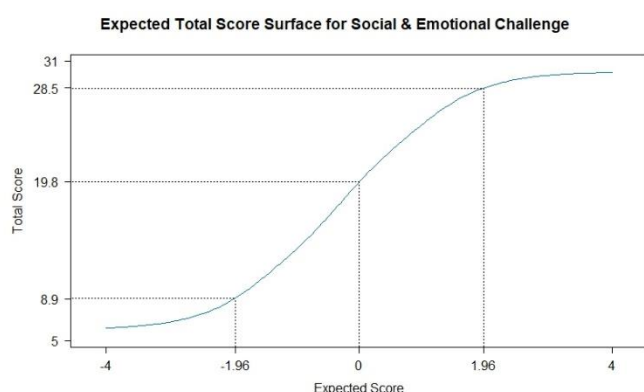
Note. Dotted Vertical lines at $\theta = -2, 0, \text{ and } 2$.

Figure 31 displays the expected scores derived from the test characteristic curves, plotted against the latent continuum (SE challenge). Converting latent attribute scores to sum scores helps in understanding the distribution and tendencies of responses using the original metrics of the SE challenge scale. Given that the SE challenge scale comprises six items, each with a five-point Likert scale, the range of expected scores falls between 6 and 30. The results indicated that

participants scoring below -1.96 on the latent continuum had expected scores of 8.9 or lower, suggesting they were most likely to select response options 1 or 2 on the Likert scale. By choosing ± 1.96 , which represents two standard deviations in the standard normal distribution, we can infer that approximately 95% of randomly selected participants would score between 8.9 and 28.5 with an average of 19.8.

Figure 31

Expected Total Score Graph for the SE Challenges Scale



IRT Analyses of the General Academic Challenges Scale (AC-g)

The distribution frequencies of items in each scale were reported in prior sections. The most noticeable differences among frequency distributions across items within each scale were observed in the GTM and COG scale. Items GTM3, COG 3, and COG5 had different distributions with the other items in their corresponding scales.

We also compared the frequency distributions of all 31 items to better understand their linear associations. The distribution frequencies for items varied across scales as well. Generally, students perceived *cognitive* and *metacognitive* challenges as less challenging than the other challenge areas as indicated by the higher rate of reporting disagreement in response to these challenges. On average, about 40% of respondents rated these challenges above the midpoint of the response scale (*neutral*). *Initiating and sustaining engagement* was the most challenging area as indicated by the highest rates of agreements with the ISE items. Put differently, ISE was

shown to have the most left skewed distribution, indicating a trend where, on average, about 63% of respondents rated its items above the midpoint of the response scale.

Examining the highest and lowest response options across items showed that about 4.6% to 19.9% of respondents selected *strongly disagree* across items (mean=9%; sd=3.1%) while about 5.7% to 26.2% selected *strongly agree* (mean=13.3%, sd=6). The most frequently selected response option across items was *agree* (mean=33.6%, sd=4.9%), except for items COG 3, COG4, and MET2, for which the most frequently selected response option was *disagree*.

Items exhibited a varied range of discrimination values (*alpha*) ranging from .81 to 2.9 (Table 14). According to Baker and Kim (2017), the discrimination values, between 0.65 to 1.34 can be interpreted as moderate, between 1.35 to 1.69 as high, and above 1.7 as very high discrimination estimates. Higher discrimination values indicated that the item was more effective in differentiating between individuals with varying levels of the perceived AC-g challenge.

The estimated thresholds (*beta*) varied from -3 to 2.3 across items. For instance, the threshold values ranged from 1.3 to 2.3 for the response *strongly agree*, and from -3 to 1.6 for the response option *disagree*. To illustrate the threshold parameters using the first item, struggling with *believing I can do my work* (MOT1, Table 14), an average respondent with a latent score of -1.93 had a 50% probability of choosing response 2 (*disagree*) or higher versus choosing *strongly disagree*. An average respondent with a latent score of -0.49 had an equal (50%) chance of selecting response option 3 (*neutral*) or higher, as opposed to choosing a lower response category (1 or 2). An average respondent with a latent score of 0.31 had an equal (50%) chance of selecting response 4 (*agree*) versus choosing a lower response category (1, 2, or 3). Finally, an individual with a latent score of 1.6 was equally likely to choose response 5 (*strongly agree*) as they were to choose any lower response (1, 2, 3, or 4), reflecting a 50% probability in either case.

Given that a latent score can be interpreted as akin to a z-score, which is normally distributed across a population with a mean of zero, if an individual's latent score is above 1.6,

they are significantly above the average level for the AC-g challenge being measured. In the context of survey measuring general academic challenges over the past two weeks, such a person would be likely to select the highest response option, *strongly agree*, when asked about the extent to which they struggled with believing they can do their work.

Table 14

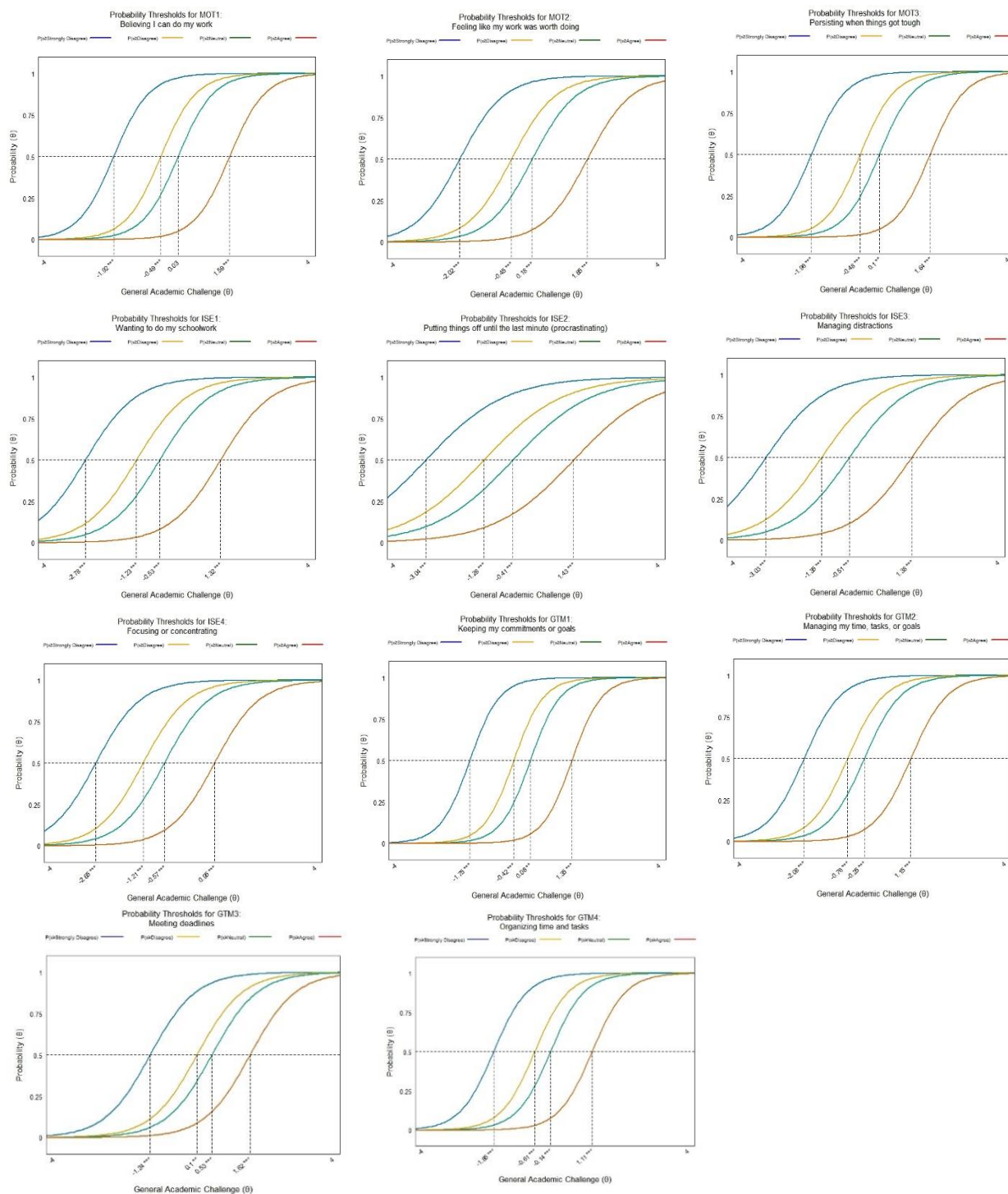
Item Response Analysis of the AC-g Scale

	α	β_1	β_2	β_3	β_4
MOT1					
par	1.871	-1.925	-0.494	0.031	1.594
SE	0.063	0.065	0.035	0.031	0.056
MOT2					
par	1.511	-2.016	-0.449	0.181	1.847
SE	0.055	0.075	0.038	0.035	0.07
MOT3					
par	1.918	-1.96	-0.48	0.099	1.636
SE	0.063	0.066	0.034	0.031	0.057
ISE1					
par	1.303	-2.782	-1.232	-0.53	1.318
SE	0.051	0.11	0.058	0.042	0.059
ISE2					
par	0.84	-3.041	-1.283	-0.407	1.433
SE	0.043	0.16	0.081	0.057	0.088
ISE3					
par	1.149	-3.035	-1.353	-0.51	1.376
SE	0.048	0.13	0.066	0.046	0.067
ISE4					
par	1.488	-2.649	-1.214	-0.572	0.946
SE	0.055	0.099	0.053	0.04	0.046
GTM1					
par	2.226	-1.753	-0.417	0.085	1.347
SE	0.072	0.057	0.031	0.029	0.046
GTM2					
par	1.802	-2.078	-0.762	-0.254	1.152
SE	0.061	0.072	0.039	0.033	0.046
GTM3					
par	1.533	-1.235	0.104	0.532	1.624
SE	0.057	0.053	0.035	0.038	0.064
GTM4					
par	1.974	-1.85	-0.613	-0.137	1.114
SE	0.066	0.063	0.035	0.031	0.043
COG1					
par	2.03	-1.829	-0.389	0.296	1.692
SE	0.066	0.061	0.032	0.031	0.057

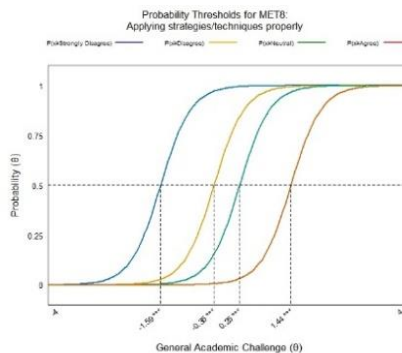
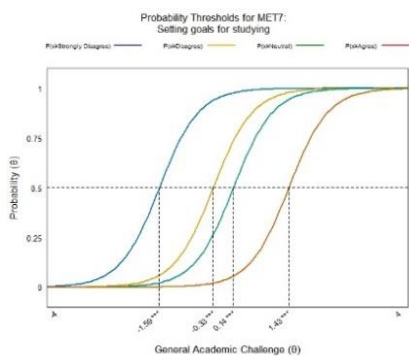
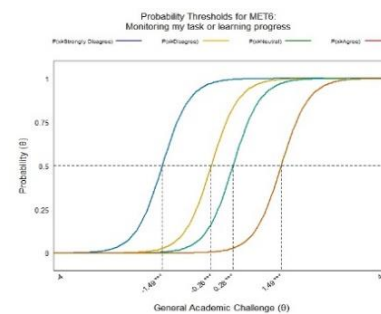
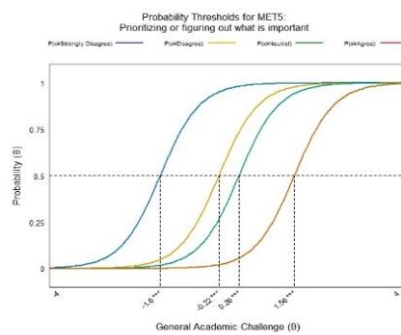
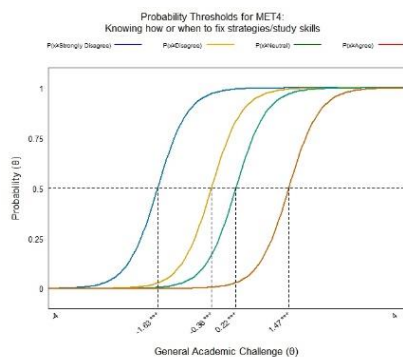
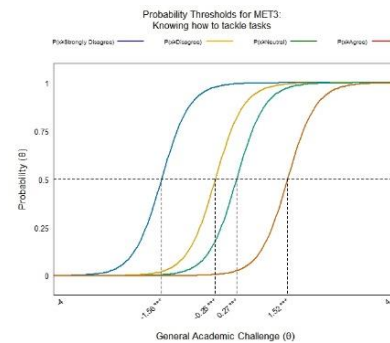
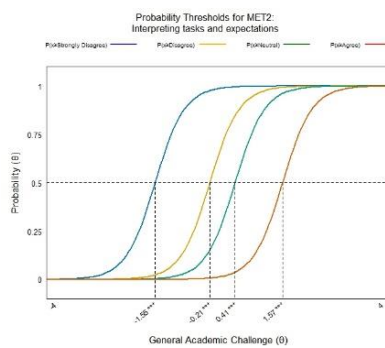
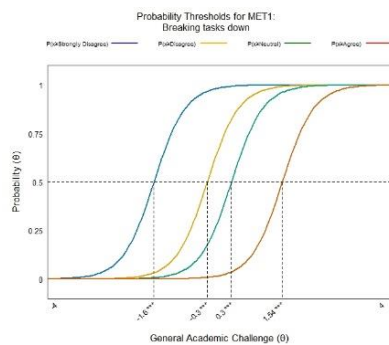
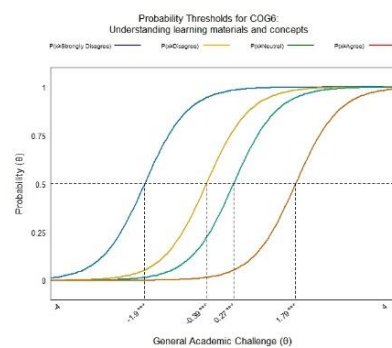
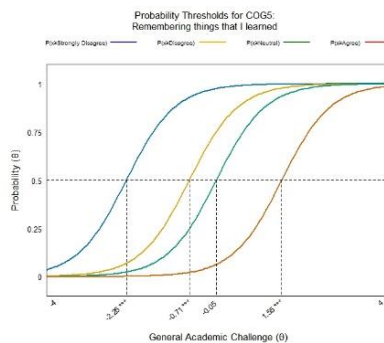
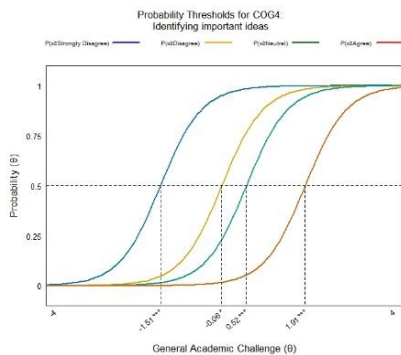
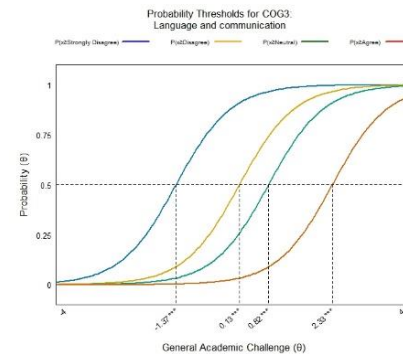
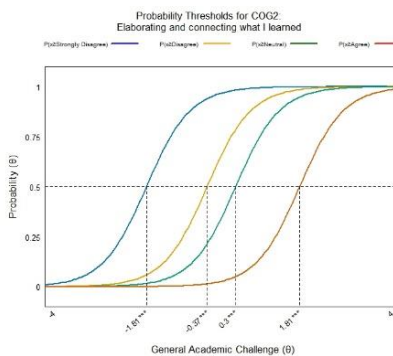
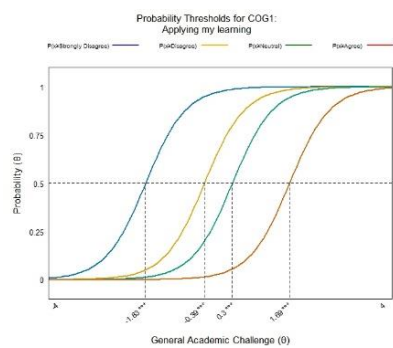
	a	b1	b2	b3	b4
COG2					
par	1.928	-1.809	-0.374	0.301	1.815
SE	0.064	0.061	0.033	0.032	0.062
COG3					
par	1.533	-1.369	0.13	0.82	2.33
SE	0.056	0.056	0.035	0.043	0.088
COG4					
par	2.056	-1.509	-0.06	0.521	1.908
SE	0.067	0.052	0.03	0.033	0.064
COG5					
par	1.677	-2.254	-0.713	-0.052	1.548
SE	0.058	0.079	0.04	0.033	0.057
COG6					
par	1.888	-1.903	-0.393	0.273	1.792
SE	0.063	0.064	0.033	0.032	0.062
SE1					
par	1.538	-2.161	-0.645	0.047	1.588
SE	0.055	0.079	0.041	0.035	0.061
SE2					
par	0.814	-3.037	-1.01	-0.045	1.876
SE	0.043	0.163	0.073	0.053	0.109
SE3					
par	1.208	-2.286	-0.683	0.171	1.671
SE	0.049	0.095	0.048	0.04	0.074
SE4					
par	1.471	-2.169	-0.699	0.007	1.455
SE	0.054	0.082	0.043	0.035	0.059
SE5					
par	1.369	-1.913	-0.397	0.382	1.89
SE	0.052	0.075	0.04	0.039	0.075
SE6					
par	1.476	-2.139	-0.789	-0.18	1.252
SE	0.054	0.08	0.044	0.03	6 0.054
MET1					
par	2.6	-1.6	-0.296	0.297	1.542
SE	0.081	0.05	0.029	0.028	0.048
MET2					
par	2.778	-1.549	-0.211	0.409	1.57
SE	0.087	0.048	0.028	0.028	0.049
MET3					
par	2.906	-1.563	-0.247	0.273	1.523
SE	0.091	0.048	0.027	0.027	0.047
MET4					
par	2.789	-1.626	-0.355	0.223	1.47
SE	0.087	0.05	0.029	0.027	0.046
MET5	a	b1	b2	b3	b4
par	2.15	1.597	-0.215	0.258	1.562
SE	0.07	0.054	0.03	0.03	0.052
MET6					
par	2.935	-1.486	-0.259	0.284	1.491
SE	0.091	0.046	0.027	0.027	0.046
MET7					
par	2.184	-1.588	-0.334	0.136	1.425
SE	0.071	0.053	0.031	0.029	0.048
MET8					
par	2.852	-1.589	-0.35	0.245	1.436
SE	0.088	0.049	0.028	0.027	0.044

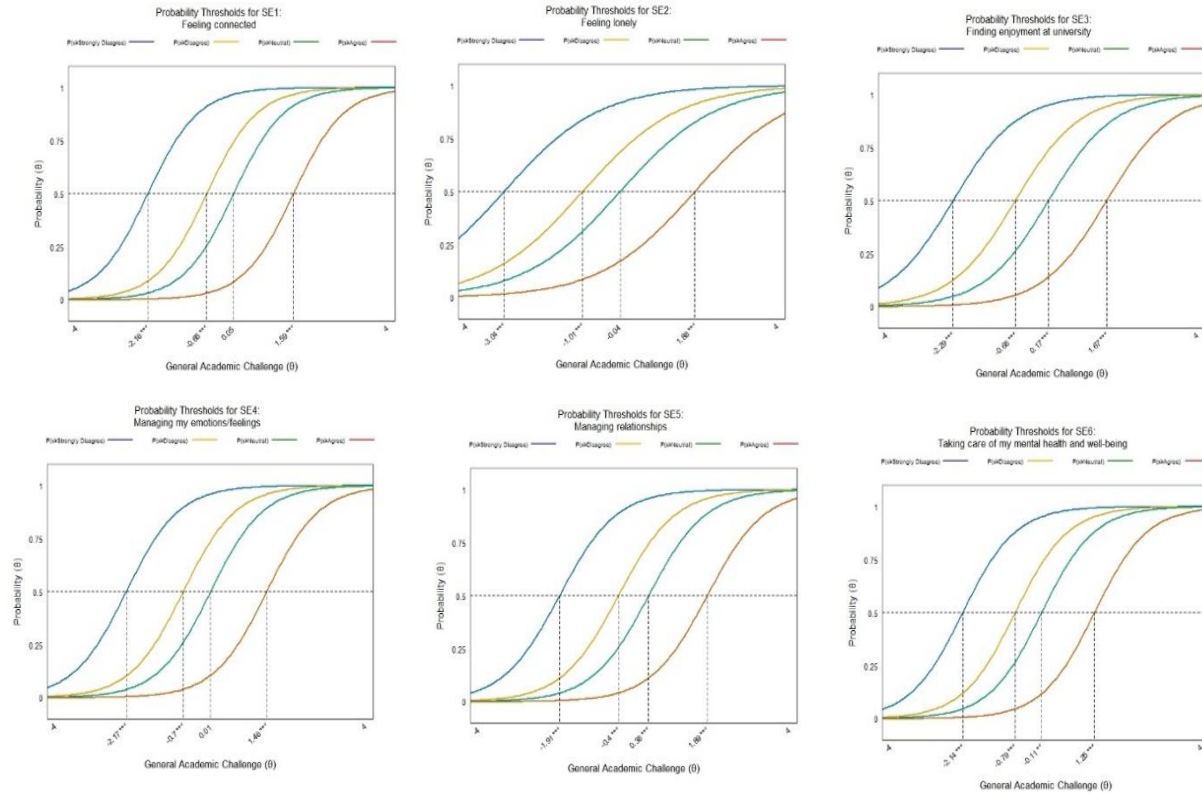
The differences between threshold parameters for consecutive response categories varied within each item (Figure 32, see also supplementary materials U). For instance, for the first item,

Figure 32
Boundary Characteristic Curves for each Item of the AC-g Scale



the distance between the thresholds for *strongly disagree-disagree* (b1-b2) and *disagree-neutral* (b2-b3) was 1.4, while the distance between thresholds for *disagree-neutral* (b2-b3) and *neutral-agree* (b3-b4) was substantially smaller, at 0.5. Across the items, ISE2, ISE3, and SE2 were shown to have larger thresholds distances compared to the other items. Consequently, an individual's total score on the AC-g challenge scale might differ from scores obtained using Classical Test Theory (CTT), where the differences between item scores are simply summed as they are assumed to be equally spaced (see Willse, 2022).



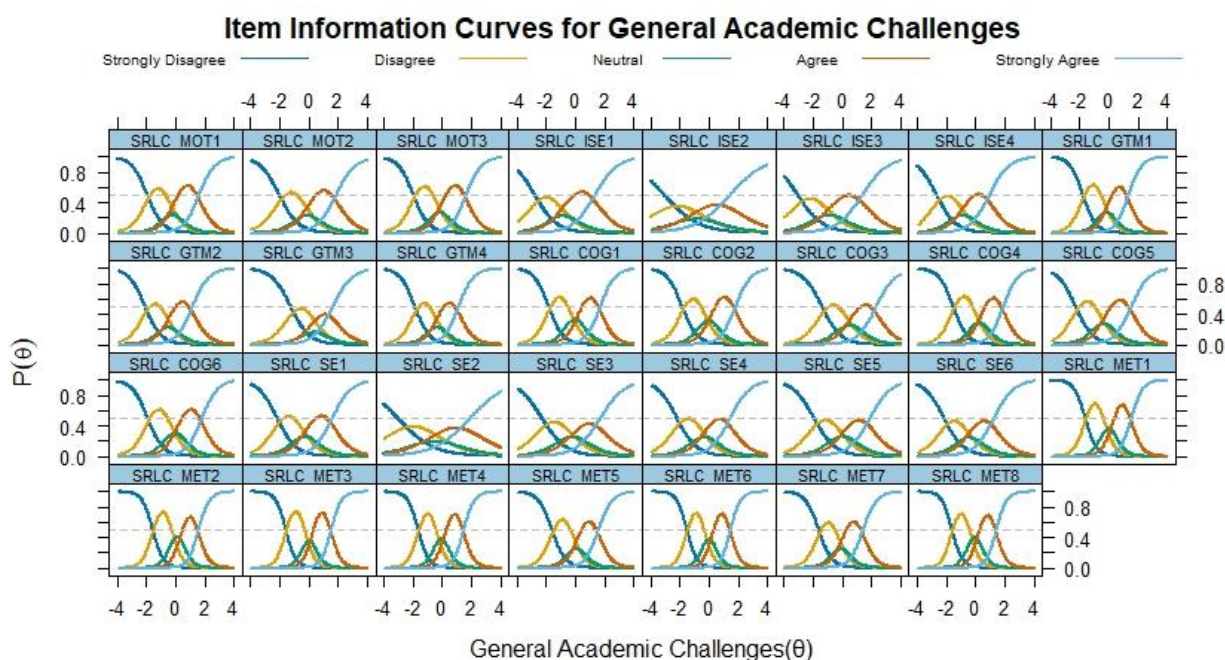


The Category Characteristic Curves (CCC) in Figure 33 demonstrated the relationship between participants' locations on the AC-g and the probability of choosing specific response options for each item. The points where two adjacent curves intersect reflect the thresholds at which endorsing the higher response option becomes equally probable. For example, for the first item, respondents with a latent score below -1.9 were more likely to respond with *strongly disagree*, while those with scores between -1.9 and -0.49 were most likely to choose *disagree*, and so forth. Additionally, these plots indicate the probability of selecting other options within this score range; for instance, for MOT1, the likelihood of selecting response option *strongly agree* was close to zero for scores within this range, while the probability of opting for *agree* was non-zero. An obvious takeaway from examining the CCCs showed that ISE2 and SE2 had significantly lower discrimination compared to the other items.

Observing the latent score range where participants were inclined to select *neutral* as the midpoint of the Likert scale showed that within this restricted range—where the likelihood of choosing *neutral* is at its peak, the probabilities of selecting response options *disagree* and *agree* exceeded that of option *neutral* in most items. The exceptions included MET1, MET2, MET3, MET4, MET6, MET8, and COG1. This observation indicated lower measurement precision at the midpoint of the Likert scale compared to the other response categories.

Figure 33

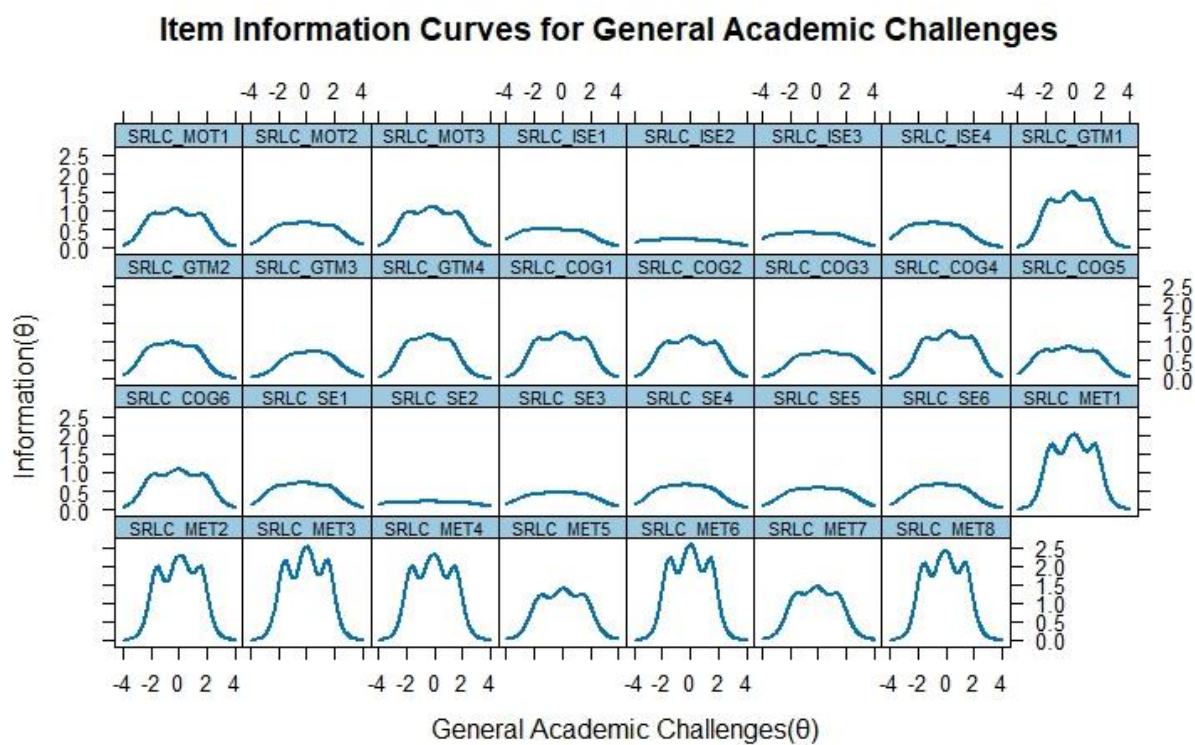
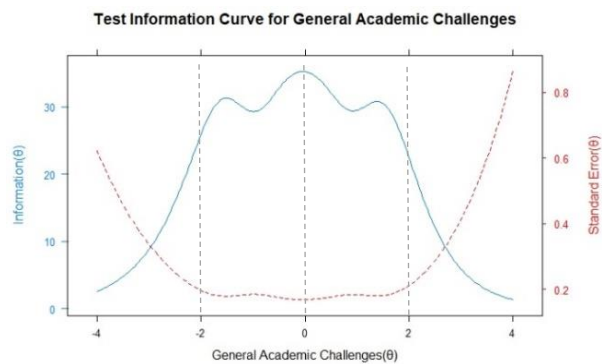
Category Characteristic Curves for each Item of the AC-g Scale



I also examined the item information function for each item (see Figure 34). The information function integrated the characteristics of threshold and discrimination parameters to indicate how much information each item contributes at different points along the latent AC-g continuum. An item that provides high information at a certain range of the latent attribute is very effective at discriminating among respondents with differing levels of the AC-g within that range.

As shown in Figure 35, items from *metacognitive* challenges subscales exhibited the highest level of information. In contrast, items ISE2 and SE2 displayed the lowest level of information and exhibited very little sensitivity in response to changes of the latent continuum of AC-g challenge compared to the other items. Comparing these results to the ones related to subscale analyses was beneficial. For instance, items ISE 3 and ISE4 were shown to provide relatively low information for measuring AC-g. However, they provided substantial information as items of ISE subscale. Therefore, the low information observed for them in the context of AC-g should be related to the unmodeled specific factor of ISE in the current analysis not to the items themselves. However, the same comparison showed that ISE2 and SE2 had poor reliability in measuring their corresponding factors as well. This observation indicated that the measurement problem is with the items themselves.

The analysis revealed that 25 out of the 31 items on the *academic challenge* scale provided moderately high to very high information, indicating their significant contribution to accurately measuring students' levels of academic challenge across a -2 to 2 *theta* range (Table 15). These high-informative items were sensitive to variations in the general academic challenge and contributed to the reliability of the AC-g construct. In contrast, the remaining 6 items displayed low to very low information, thereby contributing less to the scale's measurement precision, and offering limited sensitivity into variations in students' general academic challenges. Notably, items within the ISE and SE subscales consistently yielded lower information for the AC-g scale, despite some providing adequate information for their specific factors.

Figure 34*Item Information Function Graphs for the AC-g Items***Figure 35***Information and Standard Error Graphs for the AC-g Scale*

Note. Dotted Vertical lines at $\Theta = -2, 0$, and 2 .

Table 15*Ac-g Items Categorized in terms of Information*

Very High Information (2-2.5)	MET6.Monitoring my task or learning progress MET3.Knowing how to tackle tasks MET8.Applying strategies/techniques properly MET4.Knowing how or when to fix strategies/study skills MET2.Interpreting tasks and expectations MET1.Breaking tasks down
High Information (1.5)	GTM1.Keeping my commitments or goals MET5.Prioritizing or figuring out what is important MET7.Setting goals for studying
Moderately High Information (1)	MOT1.Believing I can do my work MOT3.Persisting when things got tough GTM4.Organizing time and tasks GTM2.Managing my time, tasks, or goals COG1.Applying my learning COG4.Identifying important ideas COG6.Understanding learning materials and concepts COG2.Elaborating and connecting what I learned
Moderate Information (.75)	COG5.Remembering things that I learned MOT2.Feeling like my work was worth doing SE1.Feeling connected SE4.Managing my emotions/feelings SE6.Taking care of my mental health and well-being COG3.Language and communication GTM3.Meeting deadlines ISE4.Focusing or concentrating
Low (.5)	ISE1.Wanting to do my schoolwork ISE3.Managing distractions SE5.Managing relationships SE3.Finding enjoyment at university
Very low information (<.5)	ISE2.Putting things off until the last minute (procrastinating) SE2.Feeling lonely

The item information functions were aggregated to form the test information function, which was used to assess the precision of the entire scale at various levels of the AC-g. A higher value in the test information function signifies higher precision in measurement and the standard error of measurement is represented by the red dotted line in the plots. As shown in Figure 31A,

the AC-g items collectively offered substantial information for measuring general academic challenges in participants scoring between -1.5 to 1.5. This indicated that the AC-g challenge scale demonstrated relatively high reliability and low measurement error within this range. The scale provided the most information and the lowest measurement error around the average *theta* level. However, the information level decreased for *theta* values below -1.5 and above 1.5.

Figure 36 illustrates the reliability for varying levels of the AC-g challenge. As depicted, the reliability remained around 0.95 for *theta* levels ranging from -2 to 2 and around .9 for *theta* levels ranging from -3 to 3. However, reliability dropped as the *theta* values moved away from this range.

Figure 36

Conditional Reliability for the AC-g Scale

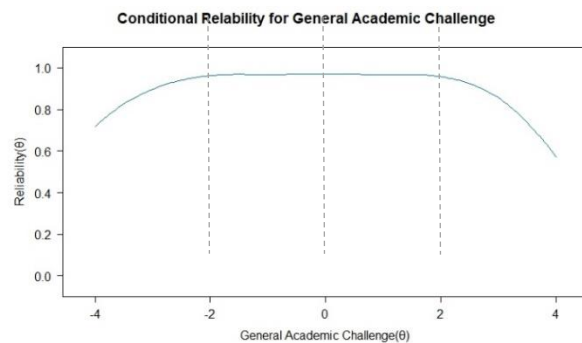
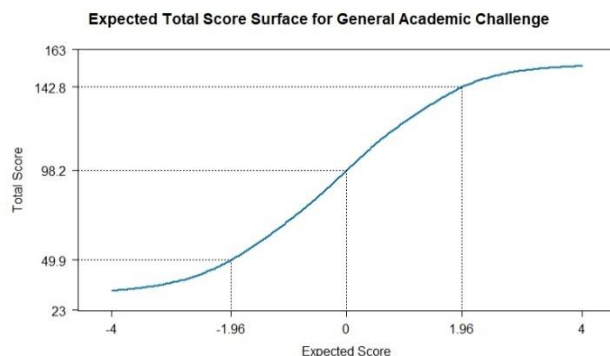


Figure 37 displays the expected scores derived from the test characteristic curves, plotted against the latent continuum (AC-g challenge). This plot can be used to convert *theta* values into corresponding total scores in the unit of the observed variables. Converting latent attribute scores to sum scores helps in understanding the distribution and tendencies of responses within the AC-g challenge context. Given that the AC-g challenge scale comprises 31 items, each with a five-point Likert scale, the range of expected scores falls between 31 and 155. The results indicated that participants scoring below -1.96 on the latent continuum have expected scores of 49.9 or lower, suggesting they were most likely to select a combination of response options 1 or 2 on the

Likert scale, while other combinations were still probable. By choosing ± 1.96 , which represents two standard deviations in the standard normal distribution, we can infer that approximately 95% of randomly selected participants would score between 49.9 and 142.8 with an average of 98.2.

Figure 37

Expected Total Score Graph for the AC-g Scale



Psychometric Adequacy of Composite Scores

Latent variable modeling, though common in psychology and education research, is not always possible in practice. For instance, in classrooms, educators often calculate sum or average scores, or in digital reports, sum/average scale scores are reported back to respondents in tabular or visual forms. Sum scores, especially when derived from multidimensional scales with various facets, are likely to have higher measurement errors (lower reliability) compared to estimated factor scores and true factors. This prompts the important question of whether a single score, derived from summing, averaging, or a factor score computed by software can accurately represent each of the subscales and the overall measure (AC-g). This section summarizes the analyses conducted to address the plausibility of calculating a composite score for each academic challenge subscale and for the AC-g scale. In this context, the term “composite” refers to both

refined composite scores (i.e., factor scores) derived from statistical analyses and *non-refined* composite scores, such as sum or average raw scores (see DiStefano et al., 2019; Grice, 2001).

Given that this section as well as the next section titled “Convergent and Criterion Validity” serve as additional sets of analyses complementing the IRT analyses, combining the rational for the analytical methods used and the reporting of findings was inevitable. This approach was adopted to ensure a cohesive presentation of the validation process.

Non-refined Composite Scores

Initially, we assessed the need to standardize scores before calculating sum/average scores. Since all items used the same response scale, examining means and variances of the item scores was essential. Means of items in each scale varied from 2.62 to 3.16, with a standard deviation of .24. Variances associated to these mean scores ranged from 1.13 to 1.35, with a standard deviation of .05. This similarity in variances across items indicated that standardization before aggregating scores is not needed. We also evaluated deviations from normality to determine the necessity of any transformations before calculating sum scores. Skewness and kurtosis values ranged from -0.70 to 0.33 and -1.21 to 0.42, respectively, suggesting no major violations of univariate normality (Byrne, 2010; Hair et al., 2010; Kline, 2015). Thus, computing sum or average scores directly from raw scores was deemed appropriate.

Average scores offer an advantage in interpretation, as they remain on the same scale as the individual items' ratings. However, without addressing missing data upfront, average scoring implicitly treats missing responses as if they were average responses. On the other hand, the scale of sum scores depends on the number of items in the scale and without addressing missing data upfront, this approach tends to conflate missing values with zeros. Within the scope of this study, considering the low rates of missing data, both scoring methods yielded comparable

results. Therefore, to facilitate easier interpretation based on the observed metrics of items, this report will concentrate on average scores moving forward.

Widaman and Revelle (2023) highlighted the importance of conceptualizing reliability as a measure of homogeneity. They argued that when a scale comprises a heterogeneous set of items, its relationship with external variables may vary across studies, potentially contributing to the so-called replicability crisis in psychology. Therefore, a closely fitting one-factor model is a key to interpreting that items form a homogeneous set, and that making a single score, such as an average is both theoretically and mathematically justified. The assumption of internal consistency is closely related to the concept of homoscedasticity of errors along the test score distribution in Classical Test Theory (CCT). This concept suggests that the variability of measurement errors should remain consistent across all levels of test score distributions, ensuring uniform reliability throughout the test. This means that regardless of whether test scores are low, medium, or high, the errors associated with these scores are expected to have a similar variance (Willse, 2022).

Widaman and Revelle (2023) also suggest that McDonald's *omega* is the preferred metric for assessing the reliability of composite scores when a unidimensional *congeneric* measurement model is appropriate for the data. Conversely, if an *essentially Tau-equivalent* measurement model is more fitting, then Cronbach's *alpha* is recommended as a reliable index for evaluating the measure's internal consistency and the overall score's reliability (see Hayes & Coutts, 2020; Revelle & Condon, 2019). In a *congeneric* model, the true scores variances (e.g., factor loadings), residuals, and covariances among variables are freely estimated. This contrasts with an *essentially Tau-equivalent* model, where true score variances are assumed equal across items, although residuals and covariances can vary (Graham, 2006). According to Table 16, the

congeneric one-factor measurement model showed an adequate fit for each subscale and for the AC-g. This finding suggested that a single dimension provided an adequate description of the relations among the items. Imposing equal factor loadings across items resulted in models with reduced fits (with the exception of the GTM and MET). However, as indicated by the fit measures, such a reduction did not result in models with a very poor fit.

Table 16

Fit Indices for Congeneric and Essentially Tau-equivalent Measurement Models

Model	X^2	df	p	CFI	TLI	RMSEA	RMSEA L	RMSEA H	SRMR
AC-g_ <i>cong</i>	10199.070	434.000	0.000	0.970	0.968	0.072	0.070	0.073	0.073
AC-g_ <i>tau</i>	7031.676	464.000	0.000	0.937	0.937	0.100	0.098	0.102	0.105
MOT_ <i>cong</i>	-	-	-	-	-	-	-	-	-
MOT_ <i>tau</i>	3.161	2.000	0.206	1.000	1.001	0.009	0.000	0.028	0.010
ISE_ <i>cong</i>	35.328	2.000	0.000	0.996	0.989	0.045	0.033	0.058	0.023
ISE_ <i>tau</i>	141.278	5.000	0.000	0.967	0.960	0.081	0.070	0.093	0.066
GTM_ <i>cong</i>	130.794	2.000	0.000	0.992	0.975	0.085	0.073	0.098	0.039
GTM_ <i>tau</i>	127.253	5.000	0.000	0.987	0.984	0.068	0.058	0.078	0.050
COG_ <i>cong</i>	526.662	9.000	0.000	0.989	0.981	0.075	0.070	0.081	0.053
COG_ <i>tau</i>	464.858	14.000	0.000	0.977	0.975	0.086	0.080	0.093	0.074
SE_ <i>cong</i>	270.279	9.000	0.000	0.990	0.984	0.060	0.054	0.066	0.043
SE_ <i>tau</i>	370.657	14.000	0.000	0.967	0.965	0.085	0.077	0.092	0.075
MET_ <i>cong</i>	617.729	20.000	0.000	0.994	0.991	0.054	0.050	0.057	0.043
MET_ <i>tau</i>	324.737	27.000	0.000	0.993	0.993	0.049	0.044	0.054	0.047

Note. Abbreviations: *cong*: Congeneric; *tau*: Essentially Tau-equivalent

Next, Cronbach's *alpha* and McDonald's *omega* were calculated for each scale and for the AC-g. These metrics produce similar results under linear item-factor relationships, low residual covariance, and a single *essentially Tau-equivalent* dimension adequately fitting the data (Widaman & Revelle, 2023). As shown in Table 17, *alpha* and *omega* estimates across the scale were either identical or very closely matched. Specifically, these estimates exceeded .9 for MET and AC-g. For the remainder of the scales, the estimates fell between .8 and .9, with the

exception of MOT and ISE, where the estimates ranged between .7 and .8. The estimates of reliability for each scale highlighted their adequate to strong internal consistencies. This observation coupled with the results of the unidimensional CFA models indicated sufficient homogeneity across items, allowing them to be represented by an overall average score for the measure.

Table 17

Frequentist Scale Reliability Statistics

	McDonald's ω	Cronbach's α	Average interitem correlation
AC-g	0.955	0.955	0.409
95% CI lower bound	0.953	0.952	0.394
95% CI upper bound	0.958	0.957	0.424
MOT	0.783	0.783	0.547
95% CI lower bound	0.767	0.768	0.524
95% CI upper bound	0.799	0.797	0.57
ISE	0.757	0.751	0.438
95% CI lower bound	0.738	0.735	0.414
95% CI upper bound	0.773	0.766	0.46
GTM			
Point estimate	0.826	0.825	0.546
95% CI lower bound	0.814	0.814	0.526
95% CI upper bound	0.838	0.836	0.566
COG	0.888	0.887	0.566
95% CI lower bound	0.881	0.88	0.549
95% CI upper bound	0.895	0.893	0.585
MET			
Point estimate	0.927	0.926	0.613
95% CI lower bound	0.922	0.922	0.596
95% CI upper bound	0.931	0.931	0.63
SE	0.843	0.841	0.47
95% CI lower bound	0.832	0.831	0.45
95% CI upper bound	0.853	0.85	0.489

In addition, as the multidimensionality was not incorporated in the unidimensional measurement model for SRL-g, we also calculated *stratified alpha*, as suggested by Rae (2007). This measure assesses the ratio of the true score variance of composite scores to the total variance of these scores, offering a more accurate reliability measure for composite scores when

subsets of items demonstrate high correlations beyond the general correlation among all items. The variability among factor loadings of a scale directly influences the *stratified alpha* making it a more conservative measure of the internal consistency for such a scale (Kamata, et al., 2003; Widhiarso & Ravand, 2014). For SRL-g, stratified alpha was .965, indicating robust internal consistency and further supporting the use of a composite average score for this scale (see supplementary materials v).

Refined Composite Scores (Factor Scores)

To further assess the quality of the non-refined composite scores for the scales (average total scores), we calculated three types of refined composite scores (factor scores) and compared their alignment with the average scores. The reason they are called *refined* is that while they do not completely account for measurement errors like a fully latent model would, they are however assumed to partially account for measurement error because they assign higher weights to observations with lower measurement error (DiStefano et al., 2019).

The three types of factor scores computed for such a comparison were: Expected A Posteriori (EAP) and EAPsum (Thissen et al., 1995; Xiong, 2023) from the IRT model, alongside Regression factor scores (Thomson, 1935) derived from factor analysis. EAP scoring estimates factor scores as the expected value of the factor's posterior distribution (Skrondal & Laake, 2001), whereas EAPsum employs a concurrent Bayesian calibration method. EAPsum transforms summed scores into scaled scores, accounting for item difficulty and discrimination, thus integrating the scale IRT-based properties with the practicality of summed scores (Thissen et al., 1995). Regression factor scores, or *exact* factor score methods, use parameters estimated from factor analysis to create linear combinations of observed variables. This type of factor scores is computed due to their widespread use (see Bartholomew, et al., 2009).

As indicated in Table 18, average scores showed high correlation with estimated factor scores (.96 to .99). Across factor scores, the EAPsum scores had the greatest correlation with average scores. McNeish and Wolf (2020) and McNeish (2023) consider a .99 correlation as indicative of equivalence between factor and average/sum scores, though Wildman and Revelle (2022) expressed more lenient views. The correlation of AC-g with EAP score was .98, suggesting that approximately 4% ($1-.98^2$) of the variability in scores differs between average scoring and factor scoring. Scatter plots (Figure 38, see also supplementary materials W) demonstrated that the deviations between average and factor scores generally occurred beyond the latent scores of -2 and 2.

Table 18

Bivariate Correlations between Average scores and Factor Scores

	EAP	EAPsum	Reg
MOT _{-avg}	0.988	0.991	0.97
ISE _{-avg}	0.956	0.988	0.937
GTM _{-avg}	0.971	0.992	0.973
COG _{-avg}	0.980	0.992	0.99
SE _{-avg}	0.973	0.978	0.97
MET _{-avg}	0.985	0.988	0.995
AC-g _{-avg}	0.979	0.992	0.993
AC-g _{-avg-means}	0.979	0.992	0.993

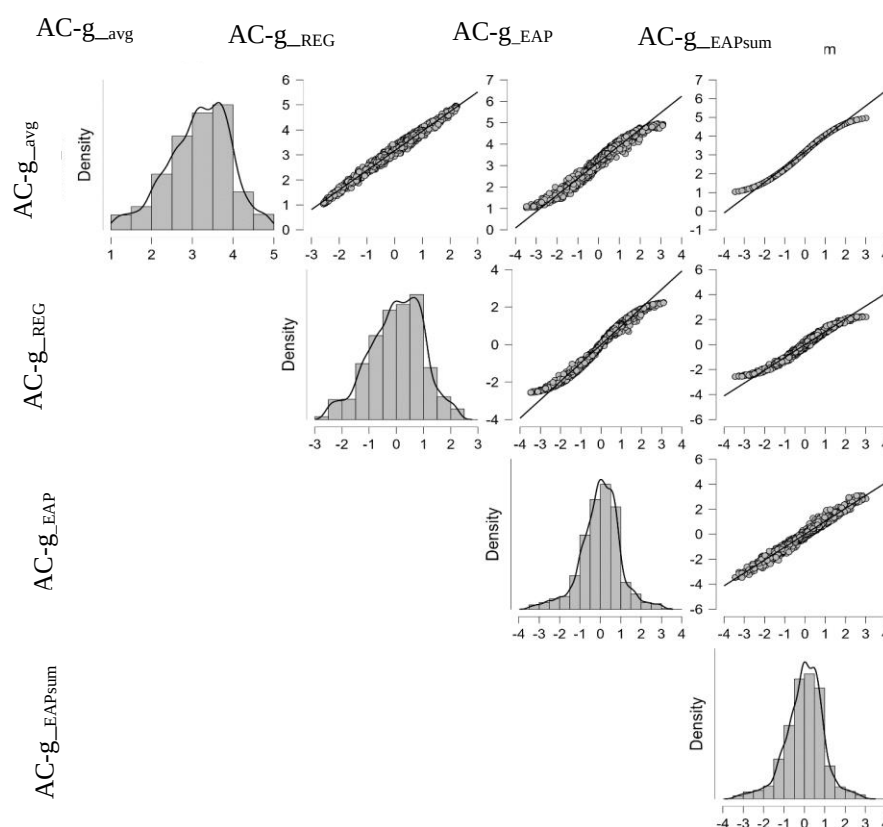
Note. AC-g_{-avg}: averaging all items; AC-g_{-avg-means}: averaging the subscales' means

Abbreviations: EAP: Expected A Posteriori; EAPsum: Expected A Posteriori for summed scores; Reg: Regression Factor Scores

We also calculated the average of the means for each subscale to examine whether scales with a greater number of items (e.g., MET) would disproportionately influence the overall score when averaging all items. To do this, we first computed the mean for each scale and then averaged these means. This approach was contrasted with directly averaging all items. The findings revealed that averaging individual items resulted in a stronger correlation with factor

Figure 38

Distributions and Linear associations between the Composite Scores



Note. Abbreviations: avg: average; EAP: Expected A Posteriori; EAPsum: Expected A Posteriori for summed scores; REG: Regression Factor Scores. Similar scatter plots for the unidimensional scales can be found in Table F in the supplementary materials.

scores (e.g., .979 vs. .967) compared to the method of averaging the means of subscales (Table 18). This observation suggested that direct item averaging provided a more accurate reflection of factor scores than aggregating subscale means.

The factor determinacy indices—which gauge the correlation between regression factor scores and true scores—varied from .91 for motivation to .98 for AC-g (Table 19). These indices highlighted the congruence between the true factors and their estimated factor scores. This index along with the minimum correlation between factor scores and the true factor revealed that factor scores for MOT, ISE, GTM, and SE align less closely with their true factors compared to other COG, MET, and AC-g. This observation will be further explored in the discussion section.

Table 19

Factor Determinacy Indices for the Scales

TRUE Factors	Correlation of (regression) scores with factors	Multiple R square of scores with factors	Minimum correlation of possible factor scores
MOT	0.91	0.83	0.66
ISE	0.92	0.85	0.69
GTM	0.94	0.88	0.76
COG	0.96	0.92	0.84
SE	0.943	0.89	0.78
MET	0.973	0.95	0.89
AC-g	0.98	0.97	0.94

In the final phase, we examined the correlation between the true AC-g factor and both the average AC-g and the AC-g factor scores. The true AC-g factor exhibited high correlations with AC-g average scores (.985), regression scores (.98), EAP scores (.996), and EAPsum scores (.997). Following a method recommended by Nicewander (2020), the reliability of the Ac-g average scores in relation to its true factor was determined to be .97. These findings supported the use of an overall average score for the entire scale, demonstrating its adequacy in capturing the construct measured by the scale.

Convergent and Criterion Validity

Convergent Validity

Examining the convergent validity of the scale began with an exploration of the Pearson bivariate correlations between factors across different scoring methods. This was followed by an assessment of factor covariances among true factors using a 6-factor solution CFA model. The factors were found to be moderately to highly correlated with each other (supplementary materials X), supporting the convergent validity of the scales.

Despite minor deviations, correlations were consistent across scoring methods, including factor scores and average scores. A comparison between the correlation matrix of the estimated scores and the one derived from the CFA model revealed that, overall, true factors exhibited stronger correlations with each other (in a 6-factor correlated factor model) than the estimated scores. This finding was expected since, unlike certain scoring methods such as the ten Berge (ten Berge, et al., 1999), the types of scoring methods employed (average scores, EAP, and EAPsum) do not preserve the covariance among factors. Also, using a correlated factor model in the presence of other construct-related sources of multidimensionality (e.g., a hierarchical factor, cross-loadings) can inflate the correlation among factors (see Rostampour et al., 2024).

Furthermore, the correlation between the estimated composite scores for the subscales and the overall AC-g scores (supplementary materials X) underscored the strong correlation of the overall AC-g score with all the subscale scores. Specifically, these associations varied from .71 to .91 for average scores ($p < .001$). This was particularly evident in the alignment of the AC-g scores with the *metacognitive* challenges scale, where the correlation exceeded .9 across different scoring methods ($p < .001$). Overall, findings emphasized the coherence of the overall scale and highlighted the significant role of *metacognitive* challenges in the broader construct measured by the AC-g scale.

Criterion Validity (Item-Level)

To evaluate the criterion validity of the scales, we explored the relationships between the subscales of academic challenges and the AC-g scale with three criterion variables: students' self-reported GPA expectations (GPA- Exp), their self-efficacy for achieving that GPA (GPA- SE), and their actual GPA at the end of the semester (GPA). The descriptive statistics for criterion variable can be found in supplementary materials B. We analyzed polychoric correlation coefficients (Drasgow,1988), as detailed in Table 20, to understand the association between individual items and these variables.

Table 20

Polychoric Correlation Estimates between individual Items and Criterion Constructs

Item	GPA		EGPA		GPA_SE	
	Coef.	<i>p</i>	Coef.	<i>p</i>	Coef.	<i>p</i>
MOT1	-0.109	***	-0.141	***	-0.312	***
MOT2	-0.111	***	-0.122	***	-0.213	***
MOT3	-0.139	***	-0.14	***	-0.239	***
ISE1	-0.061	0.025	-0.092	***	-0.14	***
ISE2	-0.215	***	-0.209	***	-0.167	***
ISE3	-0.039	0.152	-0.033	0.141	-0.121	***
ISE4	-0.043	0.117	-0.087	***	-0.158	***
GTM1	-0.154	***	-0.137	***	-0.258	***
GTM2	-0.143	***	-0.102	***	-0.217	***
GTM3	-0.246	***	-0.158	***	-0.224	***
GTM4	-0.141	***	-0.105	***	-0.235	***
COG1	-0.132	***	-0.148	***	-0.223	***
COG2	-0.074	0.006	-0.143	***	-0.23	***
COG3	-0.102	***	-0.099	***	-0.197	***
COG4	-0.129	***	-0.149	***	-0.255	***
COG5	-0.053	0.049	-0.114	***	-0.206	***
COG6	-0.105	***	-0.177	***	-0.248	***
SE1	-0.011	0.679	-0.074	0.001	-0.182	***
SE2	-0.017	0.526	-0.117	***	-0.228	***
SE3	-0.055	0.042	-0.1	***	-0.187	***
SE4	-0.004	0.88	-0.039	0.077	-0.193	***
SE5	-0.001	0.962	-0.019	0.399	-0.181	***
SE6	-0.026	0.336	-0.069	0.002	-0.213	***
MET1	-0.12	***	-0.11	***	-0.281	***
MET2	-0.121	***	-0.123	***	-0.311	***
MET3	-0.115	***	-0.145	***	-0.296	***
MET4	-0.117	***	-0.146	***	-0.307	***
MET5	-0.12	***	-0.112	***	-0.247	***
MET6	-0.142	***	-0.149	***	-0.277	***
MET7	-0.161	***	-0.15	***	-0.222	***
MET8	-0.15	***	-0.172	***	-0.283	***

Note. Abbreviations: EGPA: GPA Expectations; GPA-SE: GPA Self-Efficacy

Polychoric correlation is a technique used to estimate the correlation between two latent continuous variables from observed ordinal data. This type of correlation estimate was chosen over Pearson's due to the employment of the Diagonally Weighted Least Squares (DWLS) estimator in our SEM analyses. In other words, using polychoric correlations was needed for a valid comparison of associations derived from SEM models.

As shown in Table 21, The associations of items with *GPA Self-Efficacy* were more pronounced than those with *GPA* and *GPA expectations*. All items were negatively correlated with students' *GPA-SE* scores ($p < .001$). The strongest correlations of items with *GPA-SE* were observed with items MOT1, MET1, MET2, MET3, and MET4 ($r = -.3$), whereas the weakest correlation was with ISE3 ($r = -.12$).

The highest correlation with *GPA* was for item GTM3 ($r = -.25$, $p < .001$), followed by ISE2 ($r = -.22$, $p < .001$). Seven items, including five from the SE scale and two from the ISE scale, were found to be statistically unrelated to *GPA*.

For *GPA expectations*, the highest correlation was with ISE2 ($r = -.21$), followed by COG6 ($r = -.18$), and MET8 ($r = -.17$), all significant at $p < .001$. Three items, two from the SE scale as well as ISE3, were found to be statistically unrelated to students' *GPA expectations*.

While items from the social and emotional challenges scale were not associated with *GPA* and had a weak association with *GPA expectations* (SE1 & SE2), all of them were inversely correlated with *GPA Self-Efficacy* (average $r = -.20$, $p < .001$). Also, items such as ISE2 and SE2, which had the lowest precision of measurement across the items according to the IRT analysis, were identified as significant correlates of the criterion variables in this analysis. For instance, ISE2, which pertains to struggling with procrastination, were among the strongest

correlates of *GPA expectations* as well as actual *GPA* across the items and was associated with *GPA Self-Efficacy* as well.

Criterion Validity (Scale-Level)

To evaluate the criterion validity of the scales, the relationships of true factors, average scores, and factor scores with these variables was estimated (Table 21, see also supplementary materials Y). The associations between the true factors and these variables were evaluated using structural equation modeling with a diagonally weighted least square estimator (DWLS), while the relationships involving estimated factor and average scores were explored through polyserial correlation coefficients (Drasgow, 1988). Polyserial correlation estimates the correlation between a latent continuous variable and another observed or latent continuous variable, with at least one being represented in ordinal categories (Drasgow, 1988; Olsson, et al., 1982). We opted for this type of correlation estimate instead of Pearson's because our SEM analyses used the DWLS estimator. Therefore, polyserial correlations were necessary to ensure a valid comparison of the associations derived from the SEM models.

At the scale level, the comparison of true factors, factor scores, and average scores revealed relatively consistent correlations across different scoring methods. The standard deviation of the correlations across scoring methods ranged from .002 to .02, demonstrating consistency regardless of the scoring approach. This consistency indicated the alignment between true factors and the estimated scores for the scales.

As shown in Table 21, all polyserial correlations reflecting the associations of scales and criterion constructs were significant at $p < .001$, except for the association between the SE scale and *GPA*, which was not significant. Other scales showed a significant inverse association with *GPA*, with GTM displaying the highest correlation ($r = -.19$), followed by MOT, MET, and AC-

g ($r = -.14$). A similar pattern was observed with MOT, COG, and AC-g displaying the highest correlation ($r = -.17$) with *GPA expectation*. For *GPA Self-Efficacy*, MOT, MET, and AC-g emerged as the strongest across scales, each showing a significant correlation of $r = -.31$.

Table 21

Associations between Criterion Constructs and the Scales

Variable1	Variable2	True Factors		Average Scores		EAP Factor Scores	
		Correlation	<i>p</i>	Correlation	<i>p</i>	Correlation	<i>p</i>
MOT	GPA	-0.14	***	-0.14	***	-0.14	***
ISE	GPA	-0.09	***	-0.11	***	-0.08	0.003
GTM	GPA	-0.19	***	-0.21	***	-0.18	***
COG	GPA	-0.11	***	-0.12	***	-0.12	***
SE	GPA	-0.02	0.126	-0.02	0.42	-0.02	0.372
MET	GPA	-0.14	***	-0.16	***	-0.15	***
AC-g	GPA	-0.14	***	-0.15	***	-0.15	***
MOT	EGPA	-0.17	***	-0.16	***	-0.15	***
ISE	EGPA	-0.12	***	-0.13	***	-0.10	***
GTM	EGPA	-0.15	***	-0.15	***	-0.13	***
COG	EGPA	-0.17	***	-0.17	***	-0.16	***
SE	EGPA	-0.09	***	-0.09	***	-0.08	***
MET	EGPA	-0.16	***	-0.17	***	-0.16	***
AC-g	EGPA	-0.17	***	-0.18	***	-0.17	***
MOT	GPA-SE	-0.31	***	-0.30	***	-0.29	***
ISE	GPA-SE	-0.17	***	-0.18	***	-0.17	***
GTM	GPA-SE	-0.27	***	-0.27	***	-0.26	***
COG	GPA-SE	-0.26	***	-0.27	***	-0.26	***
SE	GPA-SE	-0.24	***	-0.25	***	-0.24	***
MET	GPA-SE	-0.31	***	-0.33	***	-0.32	***
AC-g	GPA-SE	-0.31	***	-0.33	***	-0.33	***

Note. Abbreviations: EGPA: GPA Expectations; GPA-SE: GPA Self-Efficacy

Given that students reported their *GPA Self-Efficacy* in relation to their expected GPA, polychoric partial correlation estimates were computed to quantify the relationship between academic challenges composite scores and *GPA Self-Efficacy*, controlling for their expected GPAs. These estimates, presented in Table 22, illustrate the association between *GPA Self-Efficacy* and academic challenge scales, with adjustments made to account for variations in *GPA expectation*. Even after adjusting for variations in *GPA expectations*, all scales exhibited an

inverse relationship with *GPA Self-Efficacy* ($p < .001$), with AC-g showing the strongest negative association at -0.3 and ISE displaying the weakest at -0.15.

Table 22

Partial Correlation between academic challenges composite scores and GPA confidence, controlling for GPA

Factors	Average Scores		REG		EAP		EAPsum	
	Coef.	<i>p</i>	Coef.	<i>p</i>	Coef.	<i>p</i>	Coef.	<i>p</i>
MOT	-0.27	***	-0.272	***	-0.263	***	-0.26	***
ISE	-0.153	***	-0.148	***	-0.149	***	-0.152	***
GTM	-0.248	***	-0.243	***	-0.237	***	-0.244	***
COG	-0.238	***	-0.235	***	-0.234	***	-0.239	***
SE	-0.236	***	-0.227	***	-0.227	***	-0.238	***
MET	-0.301	***	-0.304	***	-0.297	***	-0.292	***
AC-g	-0.304	***	-0.308	***	-0.303	***	-0.302	***

Note. Abbreviations: EAP: Expected A Posteriori; EAPsum: Expected A Posteriori for summed scores; Reg: Regression Factor Scores.

***: Significant at the 0.001 level ($p \leq 0.001$)

In summary, five of the six subscales significantly correlated with students' GPAs, and all scales were significantly correlated with *GPA expectations and GPA Self-Efficacy*. Notably, the SE scale was the only one without a significant association with actual GPA, yet it was linked to both *GPA expectations and GPA Self-Efficacy*. The overall scale score (Ac-g), representing general academic challenges, also showed significant associations across all examined variables. Also, all challenge scales were negatively associated with students' *GPA self-efficacy*, even after adjusting for variations in those expectations.

Discussion

In their educational journey, students frequently encounter a variety of challenges that can significantly impact their learning process. Although extensive research has been conducted on self-regulatory behaviors that foster student success, there is a notable gap in the literature concerning the specific challenges students face during their studies. These challenges, which span motivational, behavioral, cognitive, metacognitive, social, and emotional dimensions, are crucial to identify and understand from the students' perspective.

The present study sought to build upon previous work by further validating a measure of perceived academic challenges among university students using Item Response Theory (IRT). IRT analyses offer detailed information about individual items and provide insights for item refinement. In this study, each subscale was treated as a unidimensional construct and subsequently validated using IRT. Additionally, all items were combined to form a unidimensional overall construct, which was then examined through IRT.

In addition to being a measurement tool in research, the *Academic Challenge* scale is designed to serve as a self-assessment and diagnostic tool. Although the scale was initially validated as a 6-factor correlated model through confirmatory factor analysis (Hadwin et al., 2024), in practical applications such as in digital self-assessment reports and for classroom usage, sum or average scores are often employed to operationalize the scales. Consequently, the exploration of IRT parameters was complemented by an examination of the psychometric adequacy and practical utility of total scores for both subscales and the overall scale. Additionally, their associations with three criterion constructs were also investigated.

The discussion mirrors the structure used in the presentation of results, providing a concise summary of findings as well as suggestions for improving each scale individually before concluding with a *Final Remarks* section that synthesizes the findings. The suggestions for improving the scales' measurement quality presented in this study advocate for a continuous, iterative process of scale development and validation. We stress that the measurement validation process should not end once adequate psychometric indices are achieved; instead, continuous refinement between the theoretical framework and measurement models is essential.

Summary of Findings and Suggestions for Refinement of the Motivation Challenges Scale (MOT)

The analysis of the motivation challenge scale revealed several noteworthy findings. The distribution frequencies for the three items on the scale showed a consistent pattern, with each item following a similar distribution trend. This uniformity indicates that diverse respondents across multiple post-secondary institutions reacted to these items in a comparable manner. Furthermore, the scale demonstrated a balanced response tendency; it did not show a marked preference for the middle point of the Likert scale and there was enough representation of responses at both the lowest and highest ends of the scale.

In terms of item discrimination, all three items exhibited very high and uniform capabilities in differentiating between individuals who have varying levels of perceived motivation challenges. Additionally, the items displayed similar levels of information, contributing equally to the scale's measurement precision. The items collectively covered the distribution of the motivation challenge latent continuum spanning from -2 to 2 on the *theta* scale. In this range, the reliability was approximately 0.8 and the scale exhibited minimal measurement error around the average *theta* levels. According to the model, approximately 95% of randomly selected participants would score between observed scores 4.2 and 14.2, with an average score of 9.6.

As previously reviewed, the three items within this scale are designed to measure struggles with three key dimensions of motivation: self-efficacy (MOT1), value (MOT2), and persistence (MOT3). Considering the current structure of the latent model, which comprises only three items, this model is just at the threshold of statistical identification (Kline, 2015). To allow for a more nuanced understanding of the constructs being measured using latent structure

modeling, the authors suggest the addition of at least an additional item to improve the scale's capacity to discern subtle variations in different aspects of self-efficacy, value, and persistence and provide a more stable and reliable foundation for the statistical latent model. Also, despite adequate measurement quality, the reliability of this scale fell below the .8 threshold and could be enhanced by incorporating additional items.

As an example, the concept of value within the context of academic engagement is multifaceted, embodying not just an interest in the task at hand but also the perceived benefits derived from completing the task. Enhancing the scale's psychometric validity and reliability could be achieved by incorporating items that distinctly measure both these aspects of value. For example, including items like "struggled with maintaining a positive view about the benefits of studying/completing learning tasks" would address the value of the task's outcome. This, combined with the existing item (MOT2) that assesses the value attributed to the task itself, would offer a fuller evaluation of the value component. Moreover, measuring the interest in the tasks themselves through items such as "struggled with finding enough enjoyment in studying/doing a learning task" can capture intrinsic motivational factors. Adding these items would provide a balanced measure of both intrinsic and extrinsic motivational elements that influence a learner's motivation for a task (see Ryan & Deci, 2000).

Summary of Findings and Suggestions for Refinement of the Initiating and Sustaining Engagement Scale (ISE)

The analysis of the *Initiating and Sustaining Engagement Challenges Scale (ISE)* revealed several key findings regarding its item characteristics. The varied frequency distributions across items suggested distinct characteristics for each, while neither an excessive

inclination towards the midpoint nor an underrepresentation at the extreme ends of the response options was observed.

In terms of item discrimination, the analysis showed that all items, except ISE2, had very high discrimination values. This indicated their effectiveness in differentiating between varying levels of the ISE challenge. However, ISE2 stood out with only moderate discrimination, suggesting it may be less responsive to variation of the ISE challenge latent attribute.

The item information function analysis further highlighted some distinctions across items. Specifically, ISE4 provided the highest level of information, followed by ISE3 and ISE1, suggesting these items were particularly effective in measuring the ISE challenge. In contrast, ISE2 displayed the lowest level of information and showed lower sensitivity to changes in the latent continuum (*theta*), indicating a potential area for improvement. Collectively, the items demonstrated reliable information for measuring the ISE challenge for participants scoring between -2 and 1 on the *theta* scale. The reliability of the scale remained robust at around 0.8 within this range but drops to 0.6 at a *theta* value of 2, with a further decrease as the *theta* values move away from this range. Regarding response tendencies, it was estimated that approximately 95% of randomly selected participants would score between 7.2 and 19.3 on this scale, with an average score of 14.8. The high rates of agreement with ISE items and the observation that the scale's reliability was suboptimal for *theta* values above 1 indicated that participants generally found the ISE items easy to agree with, highlighting a need for item modification or adding items measuring more challenging aspects of initiating and sustaining engagement.

As discussed earlier, ISE 1 and ISE2 measure struggles with initiating the tasks while the ISE3 and ISE4 measure challenges associated with distractibility. Considering the IRT findings,

two items stood out as candidates for refinement: ISE2 and ISE1. The need for these adjustments stems from their suboptimal IRT parameters and clarity of their phrasing.

The item ISE2 (struggled with *putting things off until the last minute (procrastination)*), when coupled with response options ranging from *strongly disagree* to *strongly agree*, may present several challenges for accurate measurement. Respondents are required to interpret the item through a negative construct ("struggled with") applied to a negative behavior ("putting things off"), alongside response options that include negative choices (e.g., "Strongly disagree"). Such a semantic configuration may increase cognitive load, complicating the process for respondents to accurately reflect their experiences, and potentially affecting the item's validity.

Moreover, contrary to other items in the scales where stating disagreement implied avoiding adaptive behavior, disagreement here suggested engaging in adaptive behavior (not procrastinating). Also, the phrase "until the last minute" might predispose respondents to *agree* with the item, considering the commonality of procrastination in academic settings (Steel, 2007; Steel & Klingsieck, 2016), further complicating the item's effectiveness.

A more straightforward approach to wording this item would be "I struggled with tackling/overcoming procrastination." This revised phrasing potentially enhances the interpretability and reliability of the item. Removing "until the last minute" could shift the focus from a more deliberate and active form of procrastination commonly seen in university settings (Choi & Moran, 2009; Chun Chu & Choi, 2005) to broader procrastination tendencies including more dysfunctional ones.

The existing item ISE1, "struggled with wanting to do my schoolwork," might be better served by a slight modification to "struggled with *starting* to do my schoolwork." This change shifts the focus from a more general desire or interest in schoolwork to the specific action of

initiating the task. It addresses the element of task initiation more accurately, removing the confounding factor of direct interest measured as part of the motivation scale.

Summary of Findings and Suggestions for Refinement of the Goal and Time Management Challenges Scale (GTM)

The analysis of the Goal and Time Management challenges scale, which consists of four items, revealed some informative findings. The distribution frequencies of the scale's four items varied, indicating a different range of responses from participants across items. Importantly, the scale did not exhibit an excessive preference for the middle point on the Likert scale.

Distributions also suggested that respondents used the full spectrum of the response options, providing enough data at both the lowest and highest ends of the scale.

In terms of item discrimination, all items demonstrated high and varying levels of discrimination. This variation indicates the items' differing capacities to differentiate between individuals with varying levels of GTM challenge. Notably, item GTM2 stood out, showing a significantly higher level of information compared to the others, followed by GTM4 and GTM1. On the other hand, item GTM3 exhibited the lowest level of information and was less sensitive to changes in the latent continuum of GTM challenge than the other items.

When examining the coverage of the GTM challenge latent continuum, it was observed that items GTM2, GTM4, and GTM1 effectively spanned a range from -2 to 1.8. Reliable information was mostly obtained at an average *theta* value of 0. The scale maintained a reliability slightly above 0.8 for *theta* levels within the -2 to 1.8 range. However, a notable decrease in reliability occurred at a *theta* value of 2, dropping to 0.7, and this decline continued as *theta* values extended beyond this range. Findings showed that about 95% of randomly

selected participants are likely to score between 5.4 and 19.1 on this scale, with an average score of around 13.

When considering the item GTM3 (struggled with *meeting deadlines*) which demonstrated the lowest information among the GTM challenge items, it is important to acknowledge the contextual factors that may influence responses. The timing of data collection, particularly if it occurs in the initial weeks of a semester, can significantly impact students' responses to this item. During the early part of a semester, students might not have encountered enough instances or episodes that require them to meet deadlines, which could skew their responses. An approach to address the potential limitations of item GTM3 (struggled with *meeting deadlines*) is to incorporate a broader item. This could capture a wider range of time management issues, including self-imposed deadlines. For example, an item such as "struggled with finishing tasks on time" [or within the time frame I planned] can be more universally applicable.

Also, to obtain a more accurate assessment of students' time management challenges, it would be beneficial to include additional questions in the demographic section of the survey. These questions could ask students to rate various aspects of their academic workload, such as the phase of the semester, the overall schoolwork load for the past two weeks, the number of hours dedicated to studying, the number of courses taken, just to name a few. Gathering this information will provide valuable context and allow for a more nuanced understanding of each student's situation. Also, by incorporating these additional data points, it becomes possible to treat the varying academic demands faced by students as moderators in the analyses.

Another potential area for improvement of the scale is to reduce the content overlap between items. The scale's current items, particularly GTM2 (struggling with *managing my time*,

tasks, or goals) and GTM4 (struggling with *organizing time and tasks*), are somewhat overlapping as demonstrated by their high residual covariance.

In essence, the GTM dimension is supposed to signify challenges within the behavioral aspects of regulation. However, time management, being a complex behavior, is influenced by various facets of regulation, including metacognition and motivation (Wolters & Brady, 2020). Goal and time management challenges can occur due to struggling with planning. In this case GTM is largely a metacognitive issue, involving challenges such as poor prioritization and low task understanding (Bahena-Olivares, et al., 2024). An example of an item measuring this aspect is *struggled with planning and scheduling*, which captures the difficulties in translating intentions to plans. Another potential cause of GTM challenges is problems with translating plans into actions (behavioral aspect). Translating *intentions to plans* is the phase where goals meet plans while translating *plans into actions* is the phase where plans meet execution (Higgins, 1998; 2014; Higgins et al., 2003; Jaramillo & Spector, 2004; Jostmann & Koole, 2008; Kazén et al., 2008; Kuhl & Beckmann, 1994; Kuhl et al., 2020). This behavioural aspect can be measured using items like “*struggled with carrying out what I planned*”, reflecting difficulties in executing plans. Given that separate scales exist for measuring motivational and metacognitive challenges, this reconfiguration can create space for the behavioral dimensions crucial for effective time and task management. The items suggested in Table 23, while not exhaustive, are not meant to substitute the existing items. Instead, the table aims to showcase examples of items that can capture the behavioral challenges associated with managing goals and time.

Table 23
Suggested Items for the GTM Scale

Current Items	Alternative Items
GTM1.Keeping my commitments or goals	▪ Struggled with staying committed to my plans ▪ Struggled with carrying out what I planned ▪ Struggled with regularly working towards my goals ▪ Struggled with establishing study habits (routines) that help following my plans
GTM2.Managing my time, tasks, or goals	▪ Struggled with planning and scheduling ▪ Struggled with scheduling and allocating my time.
GTM4.Organizing time and tasks	▪ Struggled with sticking to my planned schedule ▪ Struggled with creating a schedule that I can stick to
GTM3.Meeting deadlines	▪ Struggled with finishing tasks within the time frame I planned ▪ Struggled with finishing tasks on time ▪ Struggled with finishing tasks before deadlines

Summary of Findings and Suggestions for Refinement of the Cognitive Challenges Scale (COG)

The analysis of the six items of *cognitive* challenges scale revealed several key findings. First, the distribution frequencies of these six items varied across items. Notably, there was no excessive concentration of responses at the midpoint, nor was an underrepresentation at the lowest and highest ends of the Likert response options. Regarding item discrimination, all items on the COG challenge scale displayed high to very high slopes indicating the scale's effectiveness in differentiating between individuals who possessed varying levels of the perceived *cognitive* challenge.

In terms of the amount of information provided by each item, COG2 stood out with a significantly higher level, followed by COG1, COG4, and COG6. However, items COG3 and COG5 exhibited the lowest levels of information. Their lower performance suggested a reduced

sensitivity to changes in the latent continuum of cognitive challenges, potentially limiting their effectiveness in the overall scale. Overall, the combination of items covered the distribution of the cognitive challenge latent continuum, which spanned from -2 to 2. The scale demonstrated a relatively high reliability, with a coefficient of .85, and maintained low measurement error within this range. According to the model, approximately 95% of randomly selected participants are likely to score between 7.9 and 27.9 on this scale, with an average score of 17.8.

The lower measurement precision of items COG5 and COG3 provided insights for refining the scale. Regarding item COG5 ("struggled with *remembering things I learned*"), the threshold parameters indicated that students generally found this item easier to agree with compared to other items on the scale. Despite this, this item captures a crucial aspect of cognitive regulation – the ability to recall learned information. Therefore, while it may not discriminate as effectively among those with higher perception of cognitive challenge, it remains an important element of the scale for assessing a fundamental cognitive skill. As the current item was interpreted as an easy item to agree with, an alternative item measuring a more severe challenge of this kind could be "I struggled with remembering things I learned even after reviewing them". This change is needed to better differentiate the severity of memory challenges among respondents.

Item COG3 (struggling with *language and communication*) presents a more complex scenario. This item encompasses two distinct components: struggling with language and struggling with communication. The communication aspect could be interpreted as difficulties in various cognitive processes such as expressing learned concepts, communicating understanding, sharing knowledge clearly, translating knowledge into understandable language, and discussing content coherently, etc. However, given that the sample comprised both domestic and

international students, it is possible that "struggling with language" was primarily interpreted as challenges faced while studying in a foreign language for international students. This interpretation added another layer of complexity to the item. Since the first set of interpretations (related to general communication of knowledge) might overlap with language difficulties, considering separating these components into distinct items might not be practical, as responses to potential items like "Struggled with communicating my understanding of the concepts" or "Struggled with translating my knowledge into understandable language" inherently interweave communication and language skills, reflecting the language basis of communication. However, these suggested items may be more informative for a scale measuring cognitive challenges because they address more specific and cognitively demanding challenges, which can help correct the positively skewed rates of disagreement in the original item.

Also, gathering further demographic details, like the congruence between students' native language and the instruction language, alongside self-assessed language proficiency can be advantageous. This strategy enables the scale to distinctly identify whether the challenges stem from broad cognitive and communication difficulties or are primarily linked to language proficiency. Gathering contextual information would refine the scale's precision, boosting its validity, especially in diverse international student contexts.

Additionally, exploring a potential refinement of item COG1 might also be advantageous. Originally phrased as Struggling with *applying my learning*, this item is robust in its psychometric properties. Nonetheless, adjusting its wording could improve clarity and target a more specific aspect of cognitive challenges. The suggested rephrasing includes "struggled with making learned concepts work in actual tasks" or "struggled with putting my theoretical knowledge into practice." These modifications are intended to shift from a broad, general

application of learning to a more specific and practical application. By general application, we refer to the broad and somewhat abstract idea of applying learning in various contexts, spanning both academic and real-life. On the other hand, the proposed refinement aims to more accurately capture challenges related to the practical implementation of learned knowledge, particularly within the two weeks prior to measurement in an academic setting. This could involve, for instance, the application of a mathematical concept learned in class to solve a specific engineering problem assigned during that period. Such a narrower focus could provide a clearer assessment of the students' ability to translate theoretical understanding into practice.

Summary of Findings and Suggestions for Refinement of the Metacognitive Challenges Scale (MET)

The analysis of the metacognitive challenge scale, which comprises eight items, yielded some insightful findings regarding its distribution and effectiveness. The distribution frequencies of the eight items varied across items. The distributions spanned all ranges of the latent continuum of MET challenge, with no excessive concentration of responses at the midpoint, nor an underrepresentation at the lowest or highest ends of the response scale. Regarding item discrimination, all items on the MET scale demonstrated very high slopes, and their slopes were relatively similar. This uniformity in high discrimination indicated that all items were effectively differentiating between individuals who have varying levels of perceived metacognitive challenge.

In terms of information provided by each item, MET3 and MET6 stood out with the highest levels, followed by MET4, MET2, MET8, and MET1. Items MET5 and MET7 exhibited the lowest levels of information and were less sensitive to changes in the latent continuum of metacognitive challenges compared to the other six items. Despite this, all items collectively

covered the distribution of the metacognitive challenge latent continuum (*theta*), which spanned from -2 to 2. Also, reliable information was obtained at an average MET challenge continuum, with MET3 and MET6 providing the highest information of 3.2 and MET7 the lowest at 1.5. The overall reliability of the MET challenge scale remained very high, around 0.95 for *theta* levels ranging from -1.8 to 1.8, and around 0.9 for *theta* levels extending from -2 to 2 indicating that the MET challenge scale demonstrated high reliability and low measurement error within these ranges. Finally, the model predicted that approximately 95% of randomly selected participants are likely to score between 10 and 38.3 on the scale, with an average score of 24.2.

Though MET5 and MET7 exhibited lower information relative to the other six items in the metacognitive challenge scale, their absolute psychometric properties were still robust. The following suggestions may help refining these items further.

MET5 is currently worded as struggled with *prioritizing or figuring out what is important*. This item could be rephrased as “struggled with prioritizing tasks based on their importance” or “struggled with distinguishing between high and low priority tasks”. The new phrasing avoids the ambiguity introduced by the word “or” and clarifies that the challenge lies not just in the abstract concept of prioritization but in the practical application of assessing task importance and organizing tasks accordingly. This increased specificity can enhance the item's ability to measure an individual's metacognitive skill in task prioritization.

The item MET7, struggled with *setting goals for studying*, could be refined to “struggled with setting realistic and achievable goals for studying. This modification introduces specificity to the item, better capturing the metacognitive processes involved in effective goal setting. By emphasizing the qualities of goals as realistic and achievable, the revised phrasing more accurately targets the evaluative and regulatory aspects of metacognition, enhancing the item's

focus on the strategic planning skills necessary for successful learning and academic performance. Such specificity could potentially reduce the cross-loadings between this items and items in the *goal and time management* challenges.

Additionally, while less critical, the item MET1 struggled with *breaking tasks down* could be improved to “struggled with breaking tasks down into manageable steps”. This addition further emphasizes the metacognitive skill of breaking larger tasks into smaller, more achievable segments.

Summary of Findings and Suggestions for Refinement of the Social and Emotional Challenges Scale (SE)

The analysis of the six items of the *social and emotional* challenges scale revealed varying distribution frequencies across items, with no significant inclination towards the midpoint nor underrepresentation at the highest and lowest points of the response scale. Regarding item discrimination, the items demonstrated a range from moderate to very high slopes indicating that the items were effective in differentiating between individuals with varying levels of perceived SE challenges. The only item that demonstrated poor performance relative to the others was SE2. Examining item information functions indicated that item SE4 stood out with a significantly higher level of information, followed by SE6 and SE5, suggesting these items are particularly effective in capturing variations in the latent continuum of SE challenges. In contrast, items SE1, SE3, and SE2 displayed lower levels of information, indicating a reduced sensitivity to changes in the SE challenge continuum compared to the other items.

Overall, the items collectively covered the distribution of the SE challenge latent continuum (*theta*), which spanned a range from -2 to 2. The scale was most informative and exhibits the lowest measurement error around the *theta* level of -0.2. The reliability of the scale

remained robust, around 0.9 and 0.8 for *theta* levels within the -1.8 to 1.1 and -2 to 2, respectively. According to the model's estimations, approximately 95% of randomly selected participants would score between 8.9 and 28.5 on this scale, with an average score of 19.8.

The item SE2, struggled with *feeling lonely* had a similar issue to what was observed with ISE2. In contrast to other items where disagreement suggested that a positive experience, such as feeling connected to others, was not challenging, disagreeing with this item implied that experiencing an undesirable feeling (loneliness) was not challenging. While this might not pose an issue in another context, it is noteworthy that ISE2 and SE2 were unique in having this characteristic, and both have shown the lowest precision of measurement across all items. When answering SE2, respondents must navigate a negative construct (struggled with) applied to a negative feeling (loneliness), along with options that include negative responses (e.g., disagree). This setup could heighten cognitive load, making it harder for respondents to accurately convey their experiences.

Analyzing the correlation of the SE challenge items with items in other scales showed that they were less correlated with the other scales compared to items in other scales (see supplementary materials X). This observation suggested that the SE items could be more effectively integrated into the scale if they were tailored to measure academic social and emotional challenges rather than general social and emotional challenges.

Given the title and intent of the scale, which encompasses both socialization and emotional challenges, a reorganization may be beneficial. The scale could be divided into two distinct facets: socialization challenges and emotional challenges. This separation would allow for a more focused and nuanced measurement of each aspect. The socialization dimension would assess a sense of belonging to and interpersonal connections within the academic context (Pedler,

et al., 2021; Won et al., 2018, 2019, 2022). This factor could include items such as “struggled with feeling integrated into the academic culture”, “struggled with building meaningful friendships in the academic community”, “struggled with connecting with peers/classmates/instructors at school”⁴, “struggled with establishing a supportive network at school”, “struggled with feeling like I am part of the school/university community”, and “struggled with feeling comfortable and accepted at school”. These items would collectively gauge the student’s level of integration, comfort, and connections in their academic environment.

In addition, what is labeled as an emotional challenges dimension could focus on aspects of wellbeing in the academic context. This can include struggles with managing emotions and feelings, managing stress, feeling frustrated/disappointed at university, being satisfied with one’s own academic experiences, taking care of mental health and wellbeing, etc. This dimension could be called academic wellbeing challenges. By clearly separating these two dimensions, the scale could more precisely capture the complexities of students' experiences in higher education. Identification of the various social and wellbeing challenges students face can facilitate the development of targeted interventions and support systems tailored to students' specific needs and enhance their overall academic experience.

Summary of Findings for the General Academic Challenges Scale (AC-g)

A prior study by Hadwin et al. (2024) examined validity of interpretations relating to the academic challenges Scale as a 6-factor correlated measurement model. In the current study, the

⁴ "/" was used for a concise combination of examples within a single item. However, it is crucial to separate the suggestions into distinct items to prevent double-barreling. When items merge two or more themes, respondents might concentrate on different themes within the same item, which can increase measurement errors. Separating these themes into individual items ensures clearer interpretation and more accurate data collection.

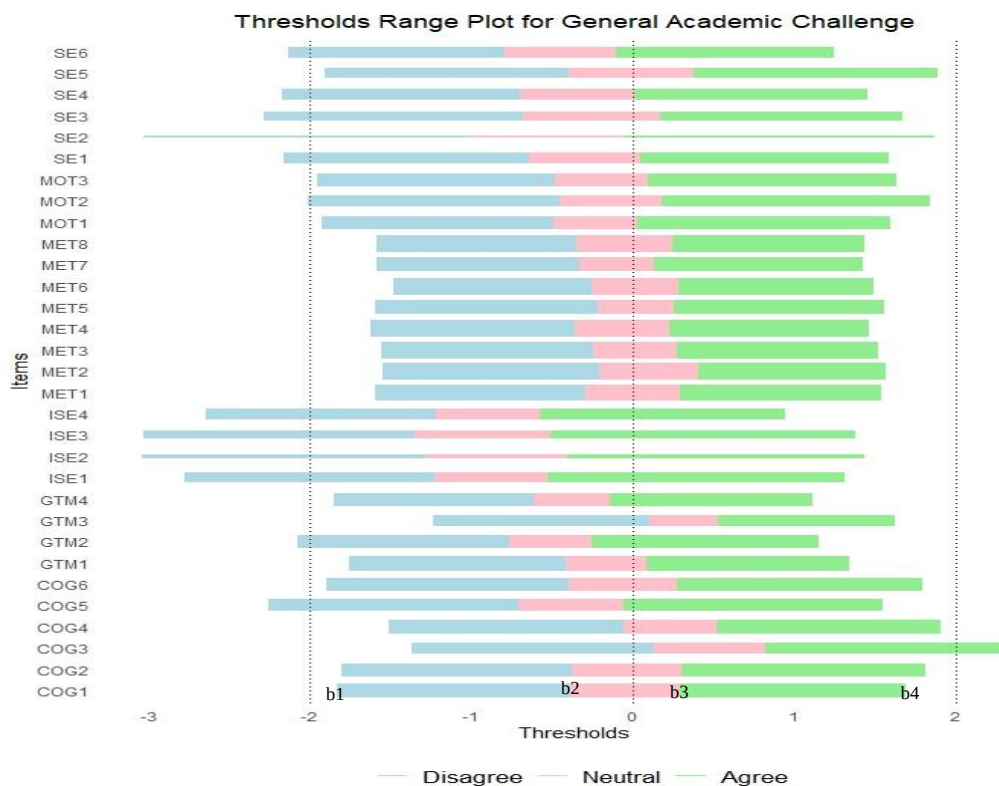
correlation between factors ranged from .54 (ISE-COG) to .85 (ISE-GTM), with the average of bivariate correlations across factors being .71 (supplementary materials X). This observation indicated a high level of shared variability among scales, suggesting the possibility of a general construct underlying the covariance structure of the items. The goodness-of-fit indices supported that a unidimensional CFA model fits the data adequately, and examination of internal consistency estimates indicated high reliability of this unidimensional scale. These observations supported that a bifactor measurement model, which accounts for both general and specific factors, would best fit the data. However, the primary objective of the study was to validate the scales as discrete constructs. Unidimensional IRT was chosen as the analytic method as it facilitated connecting the information regarding the validity of the scales to the validity of scores that can be easily computed for being used in real-world settings. As a result, while multidimensionality due to specific academic challenges factors was not considered in the AC-g unidimensional constructs, evidence of local dependency (high residual covariance) was observed for 9 pairs of items. To test whether this level of dependency skewed IRT parameters, two IRT models were compared: one including all items and another removing one item from each dependent pair. The comparison revealed that the IRT parameters were consistent across the two models. As discussed in the results section, this finding supported the stability of unidimensional IRT parameters even in the presence of certain degrees of local dependencies attributed to unmodeled multidimensionality.

The main IRT analysis of the general academic challenges scale, which comprises all 31 items, has yielded some important findings regarding its distribution and effectiveness. The distribution frequencies of the items varied across items, while there was no excessive concentration of responses at the midpoint, nor was there any data gap at the lowest or highest

ends of the response scale. Generally, students perceived cognitive and metacognitive challenges as less challenging than the other challenge areas as indicated by the higher rate of reporting disagreement in response to these challenges. *Initiating and sustaining engagement* was the most challenging area as indicated by the highest rates of agreements with the ISE items (Figure 39).

Figure 39

Thresholds Range Plot for AC-g Scale



Note. The horizontal bars represent threshold ranges for each item (illustrated using item COG1). The thickness of these bars signifies their level of discrimination, with thicker bars indicating greater discrimination. The leftward floating position of the bars suggests a high rate of student agreement with the items, while a relative position to the right indicates a high rate of disagreement with the items. Vertical dotted lines at $\theta = -2, 0$, and 2 are drawn to facilitate comparisons among the items and to illustrate the latent ranges where the items fail to measure reliably. Similar plots for each subscale can be found in the Figure E in the supplementary materials U.

Regarding item discrimination, items exhibited a varied range of discrimination values ranging from low to very high. The analysis revealed that 17 out of the 31 items on the academic challenge scale provided high and 8 items provided moderately high information indicating their significant contribution to accurately measuring students' levels of general academic challenge

across a -2 to 2 range (Table 34). This analysis also showed that across the scales items from *metacognitive challenges* subscales exhibited the highest level of information. In contrast, items ISE2 and SE2 displayed the lowest level of information and exhibited very little sensitivity in response to changes of the latent continuum of AC-g challenge compared to the other items.

The quality of the items that provided low or minimal information for the AC-g scale should be assessed against their performance for their respective specific factors. Where items demonstrated poor quality across both specific and general factors, recommendations for their refinement outlined in previous sections apply. An example was items ISE2 and SE2 which exhibited poor reliability in measuring their corresponding factors as well as in measuring AC-g. This observation indicated that the measurement problem is with the items themselves. In contrast, items ISE3 and ISE4 were shown to provide relatively low information for measuring AC-g. However, they provided substantial information as items of the ISE subscale. Therefore, their low contribution to the AC-g is likely to be related to the unmodeled specific factors in the measurement model not to the quality of items themselves.

Overall, all items collectively covered the distribution of the general academic challenge latent continuum (*theta*). The reliability remains around 0.95 for *theta* levels ranging from -2 to 2 and around .9 for *theta* levels ranging from -3 to 3. Compared to subscales, the AC-g had the highest reliability and lowest measurement error across a wider range of *theta*. Finally, the model predicted that approximately 95% of randomly selected participants are likely to score between 49.9 and 142.8 with an average of 98.2. The potential refinement of the AC-g scale, as well as the psychometric adequacy of composite scores, will be addressed as part of the next section.

Synthesis of Findings and Final Remarks

While much research focuses on self-regulatory behaviors that promote student success, there is a dearth of exploration into the specific challenges students encounter that impede learning during their studies. Viewing these challenges through a self-regulatory lens is crucial, as it is not merely about recognizing barriers but about identifying key opportunities for developing more refined and effective learning strategies. To elaborate, encountering challenges sets a stage for adaptive behavior, encouraging students to reassess ineffective strategies and adopt more successful self-regulation methods (Hadwin et al., 2019). However, when students' attempts to regulate their learning prove inadequate, especially in the face of persistent and significant challenges, potential growth opportunities can transform into impediments that hinder learning and engagement (Boekaerts, 2011; Tuominen-Soini, et al., 2012; Zimmerman & Cleary, 2009). This dual nature of challenges underscores the necessity for early identification of challenges and supporting learners to develop effective self-regulatory skills in response to those challenges.

The choices learners make to manage their studying are shaped by their perceptions of challenge experiences, among other factors. Therefore, managing one's subjective experience constitutes a crucial aspect of self-regulation. Given the subjective nature of challenge, self-report measures offer a tool for evaluating learners' perceptions of their experiences with challenges.

The current study aimed to further psychometrically examine a measure of perceived academic challenges in university students (N=2671) using item response theory (IRT). This measure, known as the *Academic Challenges* scale, is incorporated into the Self-Regulated Learning Profiles and Self-Diagnostic tool (SRL-PSD, Hadwin, 2024). It comprises 31 items

across six subscales, each evaluating the range of challenges students have faced in their studying over the last two weeks. These subscales are *motivation, initiating and sustaining engagement, goal and time management, cognitive, metacognitive, and social and emotional challenges*.

The *academic challenge* Scale is crafted with three primary objectives in mind; to be used as a) as a research measure, b) as a diagnostic and screening tool for classrooms or institutional use, and c) as a self-assessment and self-reflection tool for students. As a research instrument and a diagnostic tool in practical contexts, this scale offers a detailed profile of challenges students perceive, enriching our comprehension of their subjective experiences within educational environments. Such an insight enables the creation of targeted support strategies, moving beyond broad, generic solutions to address specific student needs. As a self-assessment mechanism, enhanced by digital reports and psychoeducational materials, it can promote a deeper self-awareness among students and empowers them to systematically identify and employ strategies to overcome their academic challenges.

Hadwin et al. (2024) previously validated the *Academic Challenges* scale as a 6-factor correlated measurement model. The present study focused on examining the scale for straightforward application in digital reports and classrooms, where complex latent statistical modeling is not available, and simple calculations are often preferred. To this end, each factor of the scale was considered as a single, unidimensional construct, and their psychometric characteristics were analyzed. The study's shift to unidimensional measurement models was designed to simplify generalizing insights obtained from item response theory to scores calculated through basic arithmetic methods, like summing or averaging.

Furthermore, the average correlation between factors at .71 suggested that, although the scales were designed to capture distinct dimensions of a broader construct, these challenge areas were not entirely independent and instead shared underlying themes. High correlations might raise concerns regarding discrimination, indicating the subscales may not be sufficiently distinct. However, within the academic challenges context, such a covariance structure could imply that the subscales may collectively represent an overarching construct of general academic challenges. Considering the interconnected nature of academic challenges, it becomes pertinent to question the feasibility of deriving a unidimensional score to represent a general factor of academic challenges. Consequently, all items were integrated into a single dimension called AC-g (general academic challenges scale).

Employing Item Response Theory (IRT), specifically through Samejima's Graded Response Model (GRM; 1997, 2016), as an alternative analytical method to triangulate the psychometric quality of the *academic challenges* scale brought several advantages. Unlike Confirmatory Factor Analysis (CFA), which assesses the overall fit of a theoretical model to the observed data, IRT focuses on the individual item level. This distinction enabled IRT to complement CFA by providing a detailed evaluation of each item within the *academic challenges* scale, thus offering a more detailed understanding of its psychometric properties.

CFA presupposes that the quality of measurement is consistent across all levels of the latent attributes. In contrast, IRT assumes that the quality of measurement may vary based on the individual's level of the latent attribute being measured (Embretson & Reise, 2013). This characteristic of IRT was particularly relevant to the scale given its diagnostic aims, as it was crucial to understand the range of academic challenges it can reliably measure. Furthermore, IRT

provides the means to evaluate the effectiveness of the response scale in distinguishing among individuals (Ostini & Nering, 2006).

The foundational assumptions of IRT, including unidimensionality, local independence, monotonicity, and Differential Item Functioning (DIF), were evaluated for all subscales and the overall academic challenges scale (AC-g). The analysis confirmed that unidimensional models were good fit for the data, indicating the appropriateness of the unidimensional approach for each scale. Furthermore, the internal consistency of the scales was examined, revealing adequate reliability for the unidimensional constructs, including the AC-g scale. The study also compared Graded Response Model (GRM) configurations with either constrained or unconstrained discrimination parameters. The findings suggested that models with unconstrained discrimination parameters were more suitable. The study aimed to understand the nuanced functioning of each item within the scale by examining key person and item parameters such as the range of latent attributes, discriminations, thresholds, information, and reliability. It also involved converting theta scores to observed expected scores to provide norm-based reference distributions.

The findings indicated the measure's adequate quality of measurement both as independent scales and as a scale for assessing general academic challenges in their present form. Nonetheless, recognizing scale development and validation as an ongoing and iterative process, recommendations for improvements were proposed prior to this general discussion.

Specifically, two items, SE2 and ISE2, were identified as problematic within both their respective subscales and the overarching academic challenges scale. Suggestions on refining the wording of these items and/or addition of new items were provided. Particularly, the *initiating and sustaining engagement* scale was identified as a frequently encountered challenge among students, as evidenced by high agreement rates and low threshold parameters. It was suggested

that this scale could be refined by incorporating items that capture more intense levels of challenges related to initiating tasks and managing attention. Similarly, the addition of items to minimize item overlap and to align the *goal and time management* scale more closely with measuring the behavioral aspects of challenges was recommended. Although the current version of the scale demonstrates sufficient psychometric validity based on IRT analysis, it was suggested that the social and emotional aspects within the social and emotional challenges scale could be more effectively measured by dividing them into two distinct dimensions.

Given that summing and averaging scores do not account for measurement error, examining the composite scores for the measure was essential. In calculations of average scores, it is assumed that all items have equal contribution to the total score, and scales with a greater number of items may disproportionately influence the overall sum scores. As a results, the psychometric validity of a model derived from analytic techniques such as CFA, which directly addresses measurement error, does not directly extend to its composite scores (McNeish & Wolf, 2020). Findings indicated that standardization or transformation of scores before aggregation was unnecessary due to the homogeneity of variances across items and the absence of significant univariate normality violations. The evidence of unidimensionality and internal consistency reinforced the suitability of composite scores for both subscales and the overall Academic Challenges scale. Notably, average scores demonstrated a very high correlation with three types of estimated factor scores in this study (.96 to .99), revealing that averaging individual items yielded a stronger correlation with factor scores (e.g., .979 vs. .967) than averaging the means of subscales. This observation suggested that direct item averaging provided a more accurate reflection of factor scores than averaging subscale means. The study also explored the congruence between true factors and their estimated factor scores. It was found that the factor

scores for *motivation* and *initiating and sustaining engagement* challenges were aligned less closely with their true factors compared to other scales, whereas *goal and time management* and *social and emotional* scales showed adequate alignment, and *cognitive, metacognitive*, and *general academic challenges* scales very strongly aligned with their true factors.

Overall, the AC-g's true factor showed a high association with its average scores (.985), underscoring the effectiveness of using an overall average score for the entire scale. However, the findings also highlighted the need for improvement, particularly in the MOT and ISE scales, where approximately 17% and 15% differences between composite scores and true factors were observed, respectively. Examining the degree to which AC-g overall scores relate to specific factors revealed a significant alignment between the overall AC-g scores and the *metacognitive* scores. This alignment persisted even after adjusting for the higher number of MET items compared to other scales, underscoring the prominent role of metacognitive challenges in explaining the covariance among other challenge areas. This finding emphasized the importance of metacognitive aspects in the academic challenges framework and suggests further investigation into this relationship.

The investigation into creating a shorter version of the measure revealed that a shorter version cannot reliably measure specific areas of academic challenge. It was advised that to achieve comprehensive coverage and ensure reliability, each scale should contain at least 4 items, with 5 being the ideal number. The Motivation (MOT) scale, in particular, was noted as needing additional items to enhance its effectiveness. In the context of the overall academic challenges scale (AC-g), although 25 out of the total 31 items were recognized for their moderately high to very high information value, the development of a shortened scale featuring key items from each subscale should be deferred. Such a decision should follow the application

of the study's recommended modifications, allowing for a revised scale that fully reflects an updated understanding of its structure and psychometric strengths.

Finally, examining the criterion validity of the scales showed significant associations between five of the six subscales and students' end of semester GPAs. All scales also showed significant correlations with *GPA expectations* and *GPA Self-Efficacy*. The *social and emotional* scale was the only scale which was not significantly associated with actual GPA, although it was related to *GPA expectations* and *GPA Self-Efficacy*. The overall scale score (AC-g), indicative of general academic challenges, also demonstrated significant inverse relationships with all criterion variables. Although the magnitudes of these associations were modest, the findings highlight the significance of these scales in relation to academic regulation and performance. The constructs assessed by these scales demonstrated meaningful connections with students' academic outcomes, emphasizing the need to account for a wide array of factors, such as expectations and self-efficacy in understanding and supporting academic success. Future studies could explore how students' perceptions of various challenges relate to their performance expectations and self-efficacy, and subsequently, their academic performance.

Future Directions

While suggestions for item-level refinement were made earlier, the scale could benefit from general improvements, both in its structure and in the analytical methods used to model it, which are detailed in this section.

General Suggestions for Improving the Scale

Response Scale. The 5-point response scale, ranging from strongly disagree to strongly agree with neutral at its midpoint, was found to be generally effective for the scales. However, this response scale for some items, such as ISE2 and SE2, were observed to potentially increase

cognitive load for respondents. Moreover, the *neutral* response option, covered a restricted range of the latent attributes and showed overlap with adjacent response options. To address this, a scale with the options *never true*, *rarely true*, *sometimes true*, *usually true*, and *always true* is suggested, alongside a prompt like "Think about the past 2 weeks when you rate how true each statement was for you. Over the past two weeks, I struggled with...". This revised response scale more accurately captures the experience of facing challenges, as opposed to the agreement/disagreement format. Its theoretical distribution is more aligned with a continuous spectrum rather than a dichotomous agreement/disagreement model with neutral as a pivot. This adjustment has already been implemented in data collection and will be analyzed in upcoming research.

Theta Range. Regarding the range of the latent attribute measured by the subscales and the AC-g, the target range was set between -2 and 2. The *academic challenges* scale is intended as a diagnostic tool for the general student population, aiming to identify variations up to two standard deviations from the mean. Capturing the remaining 5% is crucial but falls outside the intended uses of this self-report measure, potentially necessitating a different approach designed to measure those degrees of deviation. However, to reliably measure within this chosen range ($r > 0.8$), some subscales such as ISE and GTM require item revision or the addition of new items.

Including Attention Checks. Including at least two quality control questions into the survey would be advantageous to identify and exclude participants who are not adequately engaging with the survey questions. An example of such a question could be "please respond 'Never true' to this question." Additionally, employing parallel versions of at least two items and assessing the consistency of responses between them can serve as a complementary method for

attention check. Attention checks can enhance the survey's reliability by ensuring that responses are attentive and deliberate.

Item Randomization. Given the observation of higher residual covariance among items from the same scale that are sequenced consecutively, it is crucial to randomize questions throughout the entire scale. Such randomization would reduce the possibility that the observed residual covariances stem from a method effect such as item sequencing.

General Suggestions for Improving the Analytical Methods for Measuring the Scale

Simultaneous Modeling of Specific and General Factors. The global fit of the correlated factor measurement model to data, combined with the observation that a single factor explained half of the variability among items, suggests that a bifactor measurement model—considering both specific and general factors simultaneously—would be most suitable for this measure. Employing a bifactor polytomous multidimensional Item Response Theory analysis in future studies could provide further insights into the general and specific aspects of the measure (see Immekus, et al., 2019; Toland, et al., 2017).

Need for Person-Center Study of Challenges. The observed associations among the challenge areas offer valuable insights for practitioners and researchers, showing how scores from different subscales can be combined to create a comprehensive profile of students' challenges. Understanding these profiles paves the way for a deeper understanding and the development of targeted intervention strategies tailored to the diagnosis. To clarify, the categorization of learning challenges into six distinct domains—motivation, initiating and sustaining engagement, goal and time management, cognitive, metacognitive, and social and emotional—may suggest the necessity for varied intervention approaches tailored to address each specific area. For instance, students grappling with motivational issues, such as low self-

efficacy, may benefit from confidence-boosting activities and positive feedback. Similarly, individuals facing cognitive challenges, such as difficulties in connecting concepts, might require strategies focused on enhancing conceptual comprehension. However, the profile of a student's challenges typically extends beyond a single dimension, showcasing the intricate interactions between these regulatory areas. To exemplify, procrastination serves as a prime example of how challenges in academic contexts are not isolated factors but rather the result of a convergence of various factors across different regulatory dimensions. Procrastination in a student could arise from motivational issues such as low self-efficacy or not valuing the task sufficiently. Behaviorally, it might be worsened by poor study habits, such as studying in a socially distracting environment. Cognitively, it can manifest as an avoidance strategy for difficult tasks. From a metacognitive perspective, it may stem from inadequate skills in planning and setting attainable goals. From an emotion regulation viewpoint, it can be a coping response to high stress or negative emotions linked to the task (see Steel, 2007 for a meta-analysis on procrastination). Understanding these complex contributing factors and how they are related is key to a comprehensive approach to addressing these challenges.

Given this intricacy, interpretations and the subsequent psychoeducational support should examine one's academic challenges as a relational profile rather than as a set of compartmentalized factors. This relational interconnectivity should also be taken into consideration when the scales are used in research for hypothesis testing. To illustrate, consider the following questions: what does motivation challenges as an independent factor represent? what does the covariation between motivation challenges and initiating and sustaining engagement challenges represent? what does motivation challenges after controlling its covariation with ISE challenges represent? what are the meaningful ensembles of factors that can

be employed as stand-alone structural elements? Consequently, a person-centered analysis, such as latent profile analysis (see Collins & Lanza, 2009; Ding, 2018; Vermunt & Magidson, 2004), could significantly enhance our comprehension of the intraindividual interconnections among challenge areas and how these profiles correlate with factors of student success. Such an analysis could also aid in developing a categorical diagnostic scoring approach for the scale, facilitating more tailored and effective support strategies for students.

Examining Students Beliefs and Behaviours in the Context of Their Perceived

Challenges. The *Academic Challenges* scale is intended for concurrent use with the other three measures in the Self-Regulated Learning Profiles and Self-Diagnostic tool (SRL-PSD):

Perceived Importance of SRL, Self-Efficacy for SRL, and Engagement in SRL practices.

Together, these tools provide a detailed and relational profile of students' self-regulated learning.

This unified profile facilitates the interpretation of students' challenges within the broader context of their learning practices and beliefs. Understanding how these elements interrelate is crucial for promoting effective learning strategies and improving student success.

Timing of Data Collection. A plausible factor that might explain the slight variations observed in expected scores across participating university was the differences in survey administration timing. The data collection occurred within the first four weeks of the academic semesters, but the start and closing dates varied among sites. Consequently, the time windows during which participants were recruited also differed across locations. The discrepancy in timing could affect students' responses, particularly in relation to their current experiences and challenges. For instance, students who completed the survey during the second week (W2) might have reported more social and emotional challenges, such as difficulties in adjusting to new social environments. This period is typically characterized by the initial phase of university life,

where students are more likely to encounter and grapple with social integration and adaptation issues. In contrast, those who filled out the survey around the fourth week (W4) could have reported a higher frequency of cognitive and metacognitive challenges. By this time, academic pressures, such as coursework, assignments, and the need for effective study strategies, generally intensify. In other words, the differences in the patterns of challenges reported could be reflective of the evolving nature of students' experiences as the semester progresses. The temporal factors, capturing different stages of the academic term, can potentially influence the level of the challenges students report as well.

Consequently, incorporating questions that probe students' experiences, such as the number of courses they are enrolled in, the intensity of their academic workload (e.g., over the past two weeks), and the phase within the semester the data are being collected, would help to contextualize the measure. This addition would enable the tracking of academic challenges over time or across different stages of the semester, offering valuable implications for both research and practice. For instance, it is important to determine how early in the semester demotivated students can be identified by the scale, allowing for targeted interventions such as workshops or counseling to be implemented, with the aim of boosting retention rates and student success.

Context Specific Use of the Scale. Given the debate in the self-regulated learning (SRL) field about whether SRL is best measured as a domain-general or domain-specific or situation-specific (Alexander et al., 2011), adopting a domain-specific approach to using the measure could be particularly beneficial. By applying the scale to individual courses, it is possible to identify the most common challenges students face in the course, whether they are related to cognition, motivation, or other areas. Such targeted insights allow for tailored adjustments to the course, aiming to better accommodate student needs and enhance their learning experiences.

Within Person and Cross-Context Reliability. Evaluating the test-retest reliability of the measure is essential for assessing its stability both in similar and different contexts. It is important to investigate variations in perceptions of challenges across disciplines, tasks, and throughout the semester, etc. This nuanced approach to assessing stability of measure is crucial, particularly if the measure is intended for use in studies with a pretest–intervention–post-test design. By understanding these variations, we ensure that the measure accurately reflects the dynamic nature of students' perceptions and can effectively gauge the impact of interventions. Additionally, to facilitate the scale's usage in repeated assessments, developing parallel forms of the measure and establishing alternate-form reliability is essential to mitigate practice effects on the measure. This approach ensures that repeated administration of the measure over time does not lead to inflated scores due to familiarity with the content.

Computerized Adaptive Scales. Investigating the sample invariant parameters of IRT is essential for the development of computerized adaptive scales (Dodd et al., 1995). These scales optimize the assessment process by tailoring the selection of items to each respondent's previous answers, rather than administering a fixed set of items to all participants. This customization is driven by an algorithm that is both theoretically grounded and refined through empirical evidence (Boyd et al., 2011). The potential to develop a computerized adaptive scale for assessing academic challenges, especially aimed at creating personalized, self-diagnostic profiles, is promising. A long-lasting debate in the field of self-regulated learning measurement is the appropriate level of measurement granularity (Winne, 2010b). Adaptive scales offer a solution by allowing for a tailored balance between the breadth and depth of the assessment. This innovation in self-report methodologies signifies a step forward in using self-report data as a

particular mode of data within a multimodal data context, enhancing their relevance and application in contemporary research.

Need For Longitudinal Designs. The idea of modeling academic challenges as a consequence of inadequate self-regulation, which subsequently triggers a more efficient phase of self-regulation, aligns with the cyclical nature of the self-regulated learning (Hadwin et al., 2018; 2019; Järvelä et al., under review, Winne & Hadwin, 1998). Employing self-report scale alongside other types of data using intensive longitudinal methods such as experience sampling methods (see Wrzus, 2017) to investigate this hypothesis could be highly beneficial.

Triangulating with Other Data Modes. Methods of measuring SRL can be divided into two categories. The first category involves understanding students' perceptions of their learning behaviours and experiences. This can be done in real time, such as with think-aloud protocols, or retrospectively via self-report questionnaires and interviews. The second category involves observational methods, such as tracking learners' actions in real time or traces of those actions during a study episode (e.g., using digital footprints). The major underlying assumption of the self-report measures of SRL is that learners who engage in self-regulatory behaviours are aware of these behaviours and can accurately recall, aggregate, and report them. The underlying assumption for behavioural observation methods is that certain observable patterns of behaviours allow for inductively inferring the intentional and purposive use of self-regulating learning.

Self-report measures provide a means for assessing learners' perceptions of how they experience and manage their study behavior. These perceptions, however, may not always align strongly with their actual behaviors (Winne & Jamieson-Noel, 2002). This potential discrepancy can arise due to various factors, including biases inherent to self-report measures, such as recall bias (Hassan, 2005) and social desirability bias (van de Mortel & Thea, 2008). However, despite

these potential pitfalls, self-report measures remain a valuable tool for gaining access to students' mental models of the types of challenges they encounter and how they typically regulate learning in response to these challenges. No matter how much these subjective perceptions deviate from objective data, appraisal of challenges and the subsequent strategy deployment is based on these self-perceptions. In other words, a major part of self-regulation is managing one's subjective experience whose experiential quality is only accessible to the self. That is why one's perceptions can be viewed as a causal element influencing behavior construction. A more holistic view of students' academic self-regulation emerges when the subjective experiences of learners is added to the investigation of other data modalities.

Students' self-reported perceptions can help indicating whether the inductive inferences researchers draw from examining trace data accurately reflect the students' intents and perceptions. However, such a multi-modal approach is not solely for cross-validation, as divergences between students' self-reported experiences and other forms of data are equally valuable for a deeper understanding of how students manage their learning and studying. To illustrate, students' perceived challenges are influenced by factors like self-awareness and memory. Certain challenges may stand out as more salient in students' recollected memories than others. For example, students with lower metacognitive self-awareness might not fully recognize their metacognitive challenges and could mistakenly attribute these issues to other factors such as motivation. In this scenario, students could benefit from re-evaluating the reasons that caused the challenges they face, and this reconceptualization might lead to a change in their strategy use.

Limitations

Despite the valuable insights gained from this study, several limitations should be acknowledged. Addressing these limitations can guide future research to build upon and enhance the findings presented here.

COVID Context. The data for this study were gathered during the COVID-19 pandemic, specifically in the Fall of 2021 and Winter of 2022. During this period, all participating students engaged in university courses via online, off-campus education. For the majority of both students and instructors, online instruction was a relatively new mode of learning and teaching. The students were geographically dispersed across Canada and internationally, with most studying from their homes. The extent of COVID-19 restrictions on socializing varied both within Canada and around the world. Furthermore, many extracurricular activities were either canceled or faced substantial limitations. These unique circumstances should be taken into consideration when interpreting the study's results. Future studies that further validate these findings under different conditions would be beneficial.

Measurement Invariance. Differential item functioning (DIF) and differential test functioning (DTF) analyses was conducted across the five data collection sites. However, they should also be conducted with respect to other characteristics, such as year of study, gender, and socioeconomic status, among others. These analyses are crucial for demonstrating the stability of measurement parameters across different demographic groups. By identifying whether certain items or the entire test behave differently for distinct groups, researchers can ensure that the scale provides valid comparisons across diverse populations. Evaluating the measurement invariance of the scale for some of three demographic characteristics (Sex at birth, Cohorts (Fall & Winter intakes), Year of Study) is being don in another study.

Grade Point Average. GPA from the participating universities was converted into a 5-point scale to match the less detailed GPA scales of two participating universities. The distribution of GPA data across these public universities, as detailed in the supplementary materials B indicated that the sample primarily consisted of high-performing students (range=0 to 4; median=3). Therefore, generalizing the findings concerning criterion validity to student populations with different characteristics warrants further validation. Moreover, due to having a separate consent procedure for collecting GPA data, the extent of missing GPA data was noticeable (33% missing) and differed across sites, due to variations in consent processes and institutional regulations. Calculating semester GPA was also challenging, given the different frequencies of year-long courses across sites. However, when this study is compared to an earlier study by Hadwin et al. (2024), which distributed the survey towards the end of an academic semester to 476 students from a single university, the findings aligned, although the effect sizes in the prior study were marginally larger.

Reference

- Andrich, D., Marais, I., Andrich, D., & Marais, I. (2019). Violations of the assumption of independence I—multidimensionality and response dependence. In *A course in Rasch measurement theory: Measuring in the educational, social and health sciences* (pp. 173-185). https://doi.org/10.1007/978-981-13-7496-8_14
- Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R* (Vol. 969). Springer.
- Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology*, 62(3), 569-582. <https://doi.org/10.1348/000711008X365676>
- Beaumont, J., Putwain, D. W., Gallard, D., Malone, E., Marsh, H. W., & Pekrun, R. (2023). Students' emotion regulation and school-related well-being: Longitudinal models juxtaposing between- and within-person perspectives. *Journal of Educational Psychology*, 115(7), 932–950. <https://doi.org/10.1037/edu0000800>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57, 289-300. <https://doi.org/10.1111/J.2517-6161.1995.TB02031.X>
- Boekaerts, M. (2011). Emotions, emotion regulation, and self-regulation of learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 408–425). Routledge. <https://doi.org/10.4324/9780203839010.CH26>
- Boyd, A., Dodd, B., & Choi, S. (2011). Polytomous models in computerized adaptive testing. In *Handbook of polytomous item response theory models* (pp. 239–266). Routledge. <https://doi.org/10.4324/9780203861264.CH10>
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.
- Cambridge-Williams, T., Winsler, A., Kitsantas, A., & Bernard, E. (2013). University 100 orientation courses and living-learning communities boost academic retention and graduation via enhanced self-efficacy and self-regulated learning. *Journal of College Student Retention*, 15, 243–268. <https://doi.org/10.2190/CS.15.2.f>

- Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, 76, 114-140. <https://doi.org/10.1177/0013164415584576>
- Choi, J. N., & Moran, S. V. (2009). Why not procrastinate? Development and validation of a new active procrastination scale. *The Journal of Social Psychology*, 149(2), 195-212. <https://doi.org/10.3200/SOCP.149.2.195-212>
- Chun Chu, A. H., & Choi, J. N. (2005). Rethinking procrastination: Positive effects of "active" procrastination behavior on attitudes and performance. *The Journal of Social Psychology*, 145(3), 245-264. <https://doi.org/10.3200/SOCP.145.3.245-264>
- Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.
- Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2022). The Crit coefficient in Mokken scale analysis: a simulation study and an application in quality-of-life research. *Quality of Life Research*, 31(1), 49-59. <https://doi.org/10.1007/s11136-021-02924-z>
- Davis, S. K., & Hadwin, A. F. (2021). Exploring Differences in Psychological Well-Being and Self-Regulated Learning in University Student Success. *Frontline Learning Research*, 9(1), 30-43. <https://doi.org/10.14786/FLR.V9I1.581>
- De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.
- Ding, C. S. (2018). Latent profile analysis. In *Fundamentals of applied multidimensional scaling for educational and psychological research*. Springer. https://doi.org/10.1007/978-3-319-78172-3_10
- DiStefano, C., Zhu, M., & Mindrila, D. (2019). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, 14(1), 20. <https://doi.org/10.7275/DA8T-4G52>
- Distefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 425-438. <https://doi.org/10.1080/10705511.2014.915373>
- Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19(1), 5-22. <https://doi.org/10.1177/014662169501900103>

- Drasgow, F. (1988). Polychoric and polyserial correlations. In L. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 69–74). Wiley.
- Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In *The psychology of learning and motivation: Advances in research and theory* (Vol. 54, pp. 103–140). Elsevier Academic Press. <https://doi.org/10.1016/B978-0-12-385527-5.00004-8>
- Dweck, C. (2015). Carol Dweck revisits the growth mindset. *Education Week*, 35(5), 20–24.
- Dweck, C. (2016). What having a "growth mindset" actually means. *Harvard Business Review*, 13(2), 2–5.
- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, 101859. <https://doi.org/10.1016/j.cedpsych.2020.101859>
- Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, 23(1), 138. <https://doi.org/10.1037/met0000121>
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, 45(3), 225–245. <https://doi.org/10.1111/J.1745-3984.2008.00062.X>
- Fox, J., & Dusa, A. (2022). Package 'polycor': Polychoric and polyserial correlations. The Comprehensive R Archive Network.
- Gibbons, R. D., Immekus, J. C., Bock, R. D., & Gibbons, R. D. (2007). The added value of multidimensional IRT models. *Multidimensional and Hierarchical Modeling Monograph*, 1.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, 66, 930–944. <https://doi.org/10.1177/0013164406288165>
- Hadwin, A. F., & Winne, P. H. (2012). Promoting learning skills in undergraduate students. *Enhancing the quality of learning: Dispositions, instruction, and mental structures*, 201–227.

- Hadwin, A., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In *Handbook of self-regulation of learning and performance* (pp. 83-106). Routledge.
<https://doi.org/10.4324/9781315697048.CH6>
- Hadwin, A. F., Bakhtiar, A., & Miller, M. (2018). Challenges in online collaboration: Effects of scripting shared task perceptions. *International Journal of Computer-Supported Collaborative Learning*, 13(3), 301-329. <https://doi.org/10.1007/s11412-018-9279-9>
- Hadwin, A. F., Rostampour, R., Winne, P. H. (2024). *Self-regulated Learning Profile and Self-Diagnostic Tool (SRL-PSD-2021)* [Manuscript submitted for publication]. Department of Educational Psychology and Leadership Studies, University of Victoria.
- Hair, J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Educational International.
- Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., & Kirby, J. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, 95, 103208. <https://doi.org/10.1016/j.jbi.2019.103208>
- Hassan, E. (2005). Recall bias can be a threat to retrospective and prospective research designs. *The Internet Journal of Epidemiology*, 3. <https://doi.org/10.5580/2732>
- Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But.... *Communication Methods and Measures*, 14(1), 1-24.
<https://doi.org/10.1080/19312458.2020.1718629>
- Hayes, A. F., & Rockwood, N. J. (2020). Conditional process analysis: Concepts, computation, and advances in the modeling of the contingencies of mechanisms. *American Behavioral Scientist*, 64, 19-54. <https://doi.org/10.1177/0002764219859633>
- Higgins, E. T. (1998). Promotion and prevention: Regulatory focus as a motivational principle. In *Advances in experimental social psychology* (Vol. 30, pp. 1-46). Academic Press.
<https://doi.org/10.1016/S0065-2601%2808%2960381-0>
- Higgins, E. T. (2014). Beyond pleasure and pain: Value from engagement. In *Motivation and its regulation* (pp. 21-34). Psychology Press. <https://doi.org/10.4324/9781315795263-9>

- Higgins, E. T., Kruglanski, A. W., & Pierro, A. (2003). Regulatory mode: Locomotion and assessment as distinct orientations. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 35, pp. 293–344). Elsevier Academic Press.
[https://doi.org/10.1016/S0065-2601\(03\)01005-0](https://doi.org/10.1016/S0065-2601(03)01005-0)
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185. <https://doi.org/10.1007/BF02289447>
- Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. Guilford Press.
- Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, 4, 45. <https://doi.org/10.3389/feduc.2019.00045>
- Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, 63(2), 395-416. <https://doi.org/10.1348/000711009X466835>
- Jaramillo, F., & Spector, P. E. (2004). The effect of action orientation on the academic performance of undergraduate marketing majors. *Journal of Marketing Education*, 26(3), 250-260. <https://doi.org/10.1177/0273475304268780>
- Jostmann, N. B., & Koole, S. L. (2008). When persistence is futile: Action orientation and the psychology of goal disengagement. In G. B. Moskowitz & H. Grant (Eds.), *Goals* (pp. 337-361). Guilford.
- Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale scores* [Paper presentation]. Annual meeting of American Educational Research Association, Chicago, IL.
- Kang, H. A., Su, Y. H., & Chang, H. H. (2018). A note on monotonicity of item response functions for ordered polytomous item response theory models. *British Journal of Mathematical and Statistical Psychology*, 71(3), 523-535.
<https://doi.org/10.1111/bmsp.12131>
- Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, 45(4), 391-406.
<https://doi.org/10.1007/S12564-010-9082-4>
- Karimi, L., & Meyer, D. (2014). Structural equation modeling in psychology: The history, development and current challenges. *International Journal of Psychological Studies*, 6(4), 123–133. <https://doi.org/10.5539/IJPS.V6N4P123>

- Kazén, M., Kaschel, R., & Kuhl, J. (2008). Individual differences in intention initiation under demanding conditions: Interactive effects of state vs. action orientation and enactment difficulty. *Journal of Research in Personality*, 42(3), 693-715.
<https://doi.org/10.1016/J.JRP.2007.09.005>
- Kim, Y., Brady, A. C., & Wolters, C. A. (2020). College students' regulation of cognition, motivation, behavior, and context: Distinct or overlapping processes? *Learning and Individual Differences*, 80, 101872. <https://doi.org/10.1016/j.lindif.2020.101872>
- Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Koopman, L., Zijlstra, B. J., & Van der Ark, L. A. (2023). Evaluating model fit in two-level Mokken scale analysis. *Psych*, 5(3), 847-865. <https://doi.org/10.3390/psych5030056>
- Koivuniemi, M., Panadero, E., Malmberg, J., & Järvelä, S. (2017). Higher education students' learning challenges and regulatory skills in different learning situations/Desafíos de aprendizaje y habilidades de regulación en distintas situaciones de aprendizaje en estudiantes de educación superior. *Infancia y Aprendizaje*, 40(1), 19–55.
<https://doi.org/10.1080/02103702.2016.1272874>
- Kuhl, J., & Beckmann, J. (1994). *Volition and personality: Action versus state orientation*. Hogrefe.
- Kuhl, J., Quirin, M., & Koole, S. L. (2020). The functional architecture of human motivation: Personality systems interactions theory. *Advances in Motivation Science*, 7.
<https://doi.org/10.1016/bs.adms.2020.06.001>
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 529-530. <https://doi.org/10.1093/BIOMET%2F57.3.519>
- Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, 115-128. <https://www.jstor.org/stable/25051892>
- Marais, D. (2013). Local dependence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (Chapter 7). John Wiley & Sons.
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, 11, 71-101.
<https://doi.org/10.1080/15366367.2013.831680>

- McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.
- McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, 55(8), 4269-4290. <https://doi.org/10.3758/s13428-022-02016-x>
- McNeish, D. M., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52, 2287-2305. <https://doi.org/10.3758/s13428-020-01398-0>
- Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, 95(4), 728. <https://doi.org/10.1037/a0018966>
- Mindrila, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, 1(1), 60-66. <https://doi.org/10.20533/ijds.2040.2570.2010.0010>
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton.
- Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. iecProGAMMA.
- Nicewander, W. A. (2020). A perspective on the mathematical and psychometric aspects of factor indeterminacy. *Multivariate Behavioral Research*, 55(6), 825-838. <https://doi.org/10.1080/00273171.2019.1684872>
- Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, 47, 337-347. <https://doi.org/10.1007/BF02294164>
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64. <https://doi.org/10.1177/01466216000241003>
- Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289-298. <https://doi.org/10.1177/0146621603027004004>
- Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models* (Issue 144). Sage.
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Sage Publications.

- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.00422>
- Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74-98. <https://doi.org/10.1016/j.edurev.2017.08.004>
- Panadero, E., Klug, J., & Järvelä, S. (2016). Third wave of measurement in the self-regulated learning field: When measurement and intervention come hand in hand. *Scandinavian Journal of Educational Research*, 60(6), 723-735. <https://doi.org/10.1080/00313831.2015.1066436>
- Pedler, M. L., Willis, R. L., & Nieuwoudt, J. E. (2021). A sense of belonging at university: Student retention, motivation and enjoyment. *Journal of Further and Higher Education*, 46, 397-408. <https://doi.org/10.1080/0309877X.2021.1955844>
- Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In *Handbook of research on student engagement* (pp. 259-282). Springer. https://doi.org/10.1007/978-1-4614-2018-7_12
- Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review*, 18(3), 211–228. <https://doi.org/10.1007/S10648-006-9014-3>
- Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, 16, 385–407. <https://doi.org/10.1007/S10648-004-0006-X>
- Pino-Pasternak, D., Basilio, M., & Whitebread, D. (2014). Interventions and classroom contexts that promote self-regulated learning: Two intervention studies in United Kingdom primary classrooms. *Psyche*, 23(2). <https://doi.org/10.7764/PSYKHE.23.2.739>
- Revelle, W., & Revelle, M. W. (2015). Package 'psych'. The Comprehensive R Archive Network, 337(338).
- Revelle, W. (2020). An introduction to the psych package: Part II Scale construction and psychometrics. Northwestern University.
- Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research*, 54(3), 429–443. <https://doi.org/10.1080/00273171.2018.1535420>

- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25. <https://doi.org/10.18637/jss.v017.i05>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1-36.
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1), 54-67. <https://doi.org/10.1006/CEPS.1999.1020>
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer. https://doi.org/10.1007/978-1-4757-2691-6_5
- Samejima, F. (2016). Graded response models. In *Handbook of item response theory* (pp. 123–136). Chapman and Hall/CRC.
- Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In *Handbook of self-regulation of learning and performance* (2nd ed., pp. 1–15). Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9781315697048-1>
- Schunk, D. H., Meece, J. L., & Pintrich, P. R. (2013). *Motivation in education: Theory, research, and applications*. Springer.
- Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, 70(1), 137-158. <https://doi.org/10.1111/bmsp.12078>
- Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, 66, 563-575. <https://doi.org/10.1007/BF02296196>
- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1), 65. <https://doi.org/10.1037/0033-2909.133.1.65>
- Steel, P., & Klingsieck, K. B. (2016). Academic procrastination: Psychological antecedents revisited. *Australian Psychologist*, 51(1), 36-46. <https://doi.org/10.1111/AP.12173>
- Stewart, S. H., Livet, A., Arjona, R. N., Keough, M. T., Thompson, K., Krank, M., ... & Conrod, P. J. (2021). Adapting the Preventure Substance Use Prevention Program into Online

- "Univenture": Adaptation process and feasibility findings. In *CCSA's Issues of Substance 2021 conference*.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Pearson Education.
- ten Berge, J. M. F., Krijnen, W. P., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and its Applications*, 289, 311-318. <https://doi.org/10.1016/S0024-3795%2897%2910007-6>
- Tarka, P. (2018). An overview of structural equation modeling: Its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, 52, 313–354. <https://doi.org/10.1007/s11135-017-0469-8>
- Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, 19(1), 39-49. <https://doi.org/10.1177/014662169501900105>
- Thomson, G. H. (1935). The definition and measurement of "g" (general intelligence). *Journal of Educational Psychology*, 26(4), 241. <https://doi.org/10.1037/H0059873>
- Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, 41(1), 1. <https://doi.org/10.1037/h0075959>
- Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, 60, 41-63. <https://doi.org/10.1016/j.jsp.2016.11.001>
- Tuominen-Soini, H., Salmela-Aro, K., & Niemivirta, M. (2012). Achievement goal orientations and academic well-being across the transition to upper secondary education. *Learning and Individual Differences*, 22, 290-305. <https://doi.org/10.1016/J.LINDIF.2012.01.002>
- Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20(11), 1-19. <https://doi.org/10.18637/jss.v020.i11>
- Van der Ark, L. A. (2010). Getting started with Mokken scale analysis in R. Unpublished manuscript. <https://sites.google.com/a/tilburguniversity.edu/avdrark/mokken>
- Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48(5), 1-27. <https://doi.org/10.18637/jss.v048.i05>
- Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In *The Sage encyclopedia of social sciences research methods* (Vol. 2, pp. 549–553). Sage.

- Whittlesey, V., & Steiner, H. H. (2021). The strategy project: An exploration of enhancing self-regulated learning in an introductory psychology course. *InSight: A Journal of Scholarly Teaching*. <https://doi.org/10.46504/16202103wh>
- White, M. A. (2023). *Integrating wellbeing and learning in schools: Evidence-informed approaches for leaders and teachers*. Taylor & Francis.
- Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, 55(2), 788-806. <https://doi.org/10.3758/s13428-022-01849-w>
- Widhiarso, W., & Ravand, H. (2014). Estimating reliability coefficient for multidimensional measures: A pedagogical illustration. *Review of Psychology*, 21(2), 111-121.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68-81. <https://doi.org/10.1006/CEPS.1999.1015>
- Willse, J. (2022). Classical test theory. In *The SAGE encyclopedia of research design* (2nd ed., Vol. 4, pp. 198-201). SAGE Publications. <https://doi.org/10.4135/9781071812082>
- Winne, P. H. (1997). Experimenting to bootstrap self-regulated learning. *Journal of Educational Psychology*, 89, 397-410. <https://doi.org/10.1037/0022-0663.89.3.397>
- Winne, P. H. (2005). A perspective on state-of-the-art research on self-regulated learning. *Instructional Science*, 33, 559-565. <https://doi.org/10.1007/S11251-005-1280-9>
- Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist*, 41, 5-17. https://doi.org/10.1207/s15326985ep4101_3
- Winne, P. H. (2010a). Bootstrapping learner's self-regulated learning. *Psychological Test and Assessment Modeling*, 52(4), 472.
- Winne, P. H. (2010b). Improving measurement of self-regulated learning. *Educational Psychologist*, 45(4), 267-276.
- Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. In *Handbook of self-regulation of learning and performance* (pp. 15-32). <https://doi.org/10.4324/9780203839010.CH2>

- Winne, P. H. (2013). Self-regulated learning viewed from models of information processing. *Self-regulated learning and academic achievement*, 145-178. <https://doi.org/10.4324/9781410601032-9>
- Winne, P. H. (2017a). Cognition and metacognition within self-regulated learning. In *Handbook of self-regulation of learning and performance* (pp. 36–48). Routledge. <https://doi.org/10.4324/9781315697048.CH3>
- Winne, P. H. (2017b). Learning analytics for self-regulated learning. In *Handbook of learning analytics* (pp. 241–249). <https://doi.org/10.18608/HLA17.021>
- Winne, P. H. (2022). Modeling self-regulated learning as learners doing learning science: How trace data and learning analytics help develop skills for self-regulated learning. *Metacognition and Learning*, 17(3), 773–791. <https://doi.org/10.1007/s11409-022-09305-y>
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In *Metacognition in educational theory and practice* (pp. 277–304). Lawrence Erlbaum Associates Publishers. <https://doi.org/10.4324/9781410602350-19>
- Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research and applications* (pp. 297–314). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203831076-17>
- Won, S., Wolters, C. A., & Mueller, S. A. (2018). Sense of belonging and self-regulated learning: Testing achievement goals as mediators. *The Journal of Experimental Education*, 86(3), 402-418. <https://doi.org/10.1080/00220973.2016.1277337>
- Won, S., Hensley, L. C., & Wolters, C. A. (2019). Brief research report: Sense of belonging and academic help-seeking as self-regulated learning. *The Journal of Experimental Education*, 89, 112-124. <https://doi.org/10.1080/00220973.2019.1703095>
- Wrzus, C. (2017). Experience sampling methods. In V. Zeigler-Hill & T. Shackelford (Eds.), *Encyclopedia of personality and individual differences*. Springer. https://doi.org/10.1007/978-3-319-28099-8_1300-1
- Xiao, Y., Liu, H., & Hau, K. (2019). A comparison of CFA, ESEM, and BSEM in test structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 26, 665-677. <https://doi.org/10.1080/10705511.2018.1562928>

- Xiong, J., Cohen, A. S., & Xiong, X. (2023). Sequential Bayesian ability estimation applied to mixed-format item tests. *Applied Psychological Measurement*, 47(5-6), 402-419.
<https://doi.org/10.1177/01466216231201986>
- Yeager, D. S., & Dweck, C. S. (2020). What can be learned from growth mindset controversies? *American Psychologist*, 75(9), 1269. <https://doi.org/10.1037/amp0000794>
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125-145.
<https://doi.org/10.1177/014662168400800201>
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187-213.
<https://doi.org/10.1111/J.1745-3984.1993.TB00423.X>
- Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, 48(3), 135-147.
<https://doi.org/10.1080/00461520.2013.794676>
- Zimmerman, B. J., & Cleary, T. J. (2009). Motives to self-regulate learning: A social cognitive account. In *Handbook of motivation at school* (pp. 261-278). Routledge.
- Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, 53(1), 141.

Supplementary Materials for Study 2

Supplementary Material A

Academic Challenges Items

The following are the list of things that can make studying and learning at university difficult. Think about the past 2 weeks when you rate how true each statement was for you. Over the past two weeks, I struggled with...

Strongly Disagree (1) / Disagree (2) / Neutral (3) / Agree (4) / Strongly Agree (5)

MOT1	Believing I can do my work
MOT2	Feeling like my work was worth doing
MOT3	Persisting when things got tough
ISE1	Wanting to do my schoolwork
ISE2	Putting things off until the last minute (procrastinating)
ISE3	Managing distractions
ISE4	Focusing or concentrating
GTM1	Keeping my commitments or goals
GTM2	Managing my time, tasks, or goals
GTM3	Meeting deadlines
GTM4	Organizing time and tasks
COG1	Applying my learning
COG2	Elaborating and connecting what I learned
COG3	Language and communication
COG4	Identifying important ideas
COG5	Remembering things that I learned
COG6	Understanding learning materials and concepts
MET1	Breaking tasks down
MET2	Interpreting tasks and expectations
MET3	Knowing how to tackle tasks
MET4	Knowing how or when to fix strategies/study skills
MET5	Prioritizing or figuring out what is important
MET6	Monitoring my task or learning progress
MET7	Setting goals for studying
MET8	Applying strategies/techniques properly
SE1	Feeling connected
SE2	Feeling lonely
SE3	Finding enjoyment at university
SE4	Managing my emotions/feelings
SE5	Managing relationships
SE6	Taking care of my mental health and well-being

Supplementary Material B*Descriptive Statistics*

	Valid	Missing	Mean	Std.	Skewness	Std. Error	Kurtosis	Std.	Minimum	Maximum
MOT1	2669	74	3.19	1.18	-0.26	0.05	-1.04	0.10	1.00	5.00
MOT2	2667	76	3.10	1.19	-0.14	0.05	-1.08	0.10	1.00	5.00
MOT3	2668	75	3.17	1.15	-0.22	0.05	-1.02	0.10	1.00	5.00
ISE1	2670	73	3.58	1.13	-0.64	0.05	-0.48	0.10	1.00	5.00
ISE2	2669	74	3.47	1.29	-0.46	0.05	-0.97	0.10	1.00	5.00
ISE3	2669	74	3.57	1.14	-0.59	0.05	-0.55	0.10	1.00	5.00
ISE4	2667	76	3.69	1.15	-0.70	0.05	-0.42	0.10	1.00	5.00
GTM1	2670	73	3.19	1.20	-0.20	0.05	-1.07	0.10	1.00	5.00
GTM2	2670	73	3.44	1.20	-0.46	0.05	-0.87	0.10	1.00	5.00
GTM3	2671	72	2.75	1.35	0.29	0.05	-1.21	0.10	1.00	5.00
GTM4	2670	73	3.37	1.24	-0.38	0.05	-1.00	0.10	1.00	5.00
COG1	2664	79	3.06	1.14	-0.08	0.05	-0.98	0.10	1.00	5.00
COG2	2666	77	3.03	1.14	-0.09	0.05	-1.01	0.10	1.00	5.00
COG3	2664	79	2.62	1.17	0.33	0.05	-0.92	0.10	1.00	5.00
COG4	2670	73	2.81	1.15	0.17	0.05	-1.02	0.10	1.00	5.00
COG5	2671	72	3.30	1.14	-0.33	0.05	-0.89	0.10	1.00	5.00
COG6	2670	73	3.05	1.13	-0.09	0.05	-1.01	0.10	1.00	5.00
SE1	2661	82	3.26	1.18	-0.28	0.05	-0.96	0.10	1.00	5.00
SE2	2662	81	3.29	1.28	-0.26	0.05	-1.10	0.10	1.00	5.00
SE3	2665	78	3.20	1.23	-0.17	0.05	-1.05	0.10	1.00	5.00
SE4	2666	77	3.29	1.21	-0.28	0.05	-0.98	0.10	1.00	5.00
SE5	2661	82	3.02	1.22	-0.04	0.05	-1.07	0.10	1.00	5.00
SE6	2670	73	3.38	1.24	-0.37	0.05	-0.95	0.10	1.00	5.00
MET1	2669	74	3.02	1.15	-0.05	0.05	-1.00	0.10	1.00	5.00
MET2	2667	76	2.94	1.13	0.07	0.05	-0.96	0.10	1.00	5.00
MET3	2668	75	3.01	1.14	-0.04	0.05	-1.05	0.10	1.00	5.00
MET4	2665	78	3.09	1.14	-0.13	0.05	-0.99	0.10	1.00	5.00
MET5	2669	74	3.00	1.20	-0.01	0.05	-1.15	0.10	1.00	5.00
MET6	2668	75	3.01	1.16	-0.05	0.05	-1.04	0.10	1.00	5.00
MET7	2669	74	3.10	1.22	-0.13	0.05	-1.13	0.10	1.00	5.00
MET8	2667	76	3.08	1.15	-0.11	0.05	-0.98	0.10	1.00	5.00
MOT_avg	2627	0	3.15	0.98	-0.26	0.05	-0.57	0.10	1.00	5.00
ISE_avg	2627	0	3.58	0.89	-0.49	0.05	-0.17	0.10	1.00	5.00
GTM_avg	2627	0	3.19	1.01	-0.25	0.05	-0.69	0.10	1.00	5.00
COG_avg	2627	0	2.98	0.92	-0.08	0.05	-0.56	0.10	1.00	5.00
SE_avg	2627	0	3.24	0.92	-0.32	0.05	-0.40	0.10	1.00	5.00
MET_avg	2627	0	3.03	0.95	-0.12	0.05	-0.52	0.10	1.00	5.00
GPA_Conf	2721	22	3.54	0.84	-0.57	0.05	0.47	0.09	1.00	5.00
E-GPA	2718	25	3.63	0.82	-0.31	0.05	-0.03	0.09	1.00	5.00
GPA	1877	866	3.08	0.99	-1.07	0.06	0.77	0.11	0.00	4.00

Note. Abbreviations in the order they appear in the table:

Challenges: MOT: Motivation; ISE: Initiating & Sustaining Engagement; GTM: Goal & Time Management; Cog: Cognitive; MET: Metacognitive; SE: Social & Emotional; AC-g: General Academic Challenges Scale.

Avg: Average.

Criterion Constructs: GPA_Conf: GPA Confidence; E-GPA: GPA Expectations; GPA: Grade Point Average. _

Supplementary Material C
Eigenvalues derived from EFA and PCA Parallel analyses

	Factors/Components							
	1	2	3	4	5	6	7	8
General Academic Challenges								
FA Actual Data	14.68	1.65	1.10	0.72	0.54	0.26	0.19	0.10
FA Simulated Data	0.30	0.22	0.20	0.18	0.17	0.15	0.14	0.13
FA Resampled Data	0.33	0.25	0.22	0.20	0.19	0.17	0.15	0.14
PC Actual Data	15.17	2.22	1.70	1.22	1.09	0.85	0.73	0.66
PC Simulated Data	1.20	1.17	1.16	1.14	1.12	1.11	1.10	1.09
PC Resampled Data	1.23	1.20	1.18	1.16	1.14	1.13	1.11	1.10
Proportion Var (EFA)	0.48							
Average Var Extracted (CFA)	0.50							
Metacognitive Challenges								
	1	2	3	4	5	6	7	8
FA Actual Data	5.47	0.31	0.04	0.03	-0.07	-0.08	-0.10	-0.13
FA Simulated Data	0.48	0.06	0.04	0.02	0.00	-0.02	-0.04	-0.07
FA Resampled Data	0.44	0.06	0.04	0.02	0.00	-0.02	-0.04	-0.06
PC Actual Data	5.79	0.64	0.38	0.34	0.25	0.23	0.21	0.16
PC Simulated Data	1.09	1.06	1.03	1.01	0.99	0.97	0.94	0.91
PC Resampled Data	1.08	1.05	1.03	1.01	0.99	0.97	0.95	0.92
Proportion var.	0.68							
Cognitive Challenges								
	1	2	3	4	5	6		
FA Actual Data	3.81	0.23	0.14	-0.09	-0.12	-0.17		
FA Simulated Data	0.51	0.05	0.02	0.00	-0.02	-0.05		
FA Resampled Data	0.46	0.04	0.02	0.00	-0.02	-0.04		
PC Actual Data	4.16	0.68	0.46	0.29	0.25	0.16		
PC Simulated Data	1.07	1.04	1.01	0.99	0.96	0.93		
PC Resampled Data	1.06	1.03	1.01	0.99	0.97	0.94		
Proportion var.	0.63							
Social & Emotional Challenges								
	1	2	3	4	5	6		
FA Actual Data	3.24	0.26	-0.02	-0.05	-0.05	-0.13		
FA Simulated Data	0.43	0.04	0.02	0.00	-0.02	-0.05		
FA Resampled Data	0.46	0.04	0.02	0.00	-0.02	-0.04		
PC Actual Data	3.67	0.75	0.56	0.40	0.38	0.23		
PC Simulated Data	1.07	1.04	1.01	0.99	0.96	0.93		
PC Resampled Data	1.06	1.03	1.01	0.99	0.97	0.94		
Proportion var.	0.54							
Initiating & Sustaining Engagement Ch.								
	1	2	3	4				
FA Actual Data	2.11	0.09	-0.03	-0.06				
FA Simulated Data	0.61	0.02	0.00	-0.02				
FA Resampled Data	0.51	0.02	0.00	-0.02				
PC Actual Data	2.54	0.68	0.47	0.30				
PC Simulated Data	1.05	1.02	0.99	0.95				
PC Resampled Data	1.04	1.01	0.99	0.96				
Proportion var.	0.53							
Goal & Time Management Ch.								
	1	2	3	4				
FA Actual Data	2.48	0.13	-0.03	-0.10				
FA Simulated Data	0.59	0.02	0.00	-0.02				
FA Resampled Data	0.55	0.02	0.00	-0.02				
PC Actual Data	2.85	0.56	0.36	0.23				
PC Simulated Data	1.05	1.01	0.99	0.95				
PC Resampled Data	1.04	1.01	0.99	0.96				
Proportion var.	0.62							
Motivation Challenges								
	1	2	3					
FA Actual Data	1.85	0.00	0.00					
FA Simulated Data	0.50	0.00	0.00					
FA Resampled Data	0.48	0.00	0.00					
PC Actual Data	2.23	0.41	0.36					
PC Simulated Data	1.04	1.00	0.97					
PC Resampled Data	1.03	1.00	0.97					
Proportion var.	0.62							

Note. FA: Factors; PC: Principal Components

Supplementary Material D
Item-Factor Loadings for the
Unidimensional Scales

	(β)	p	R^2
MOT1	0.752	0.000	0.566
MOT2	0.705	0.000	0.497
MOT3	0.762	0.000	0.580
ISE1	0.664	0.000	0.441
ISE2	0.459	0.000	0.211
ISE3	0.736	0.000	0.542
ISE4	0.802	0.000	0.643
GTM1	0.779	0.000	0.606
GTM2	0.825	0.000	0.681
GTM3	0.619	0.000	0.383
GTM4	0.741	0.000	0.549
COG1	0.822	0.000	0.675
COG2	0.846	0.000	0.715
COG3	0.663	0.000	0.440
COG4	0.775	0.000	0.600
COG5	0.664	0.000	0.441
COG6	0.748	0.000	0.559
MET1	0.780	0.000	0.608
MET2	0.799	0.000	0.639
MET3	0.812	0.000	0.659
MET4	0.803	0.000	0.645
MET5	0.744	0.000	0.554
MET6	0.826	0.000	0.683
MET7	0.722	0.000	0.521
MET8	0.785	0.000	0.617
SE1	0.694	0.000	0.481
SE2	0.526	0.000	0.277
SE3	0.627	0.000	0.394
SE4	0.821	0.000	0.674
SE5	0.684	0.000	0.468
SE6	0.774	0.000	0.599

Supplementary Material E*Item-Factor Loadings for the AC-g Scale*

	EFA Loadings	h^2	u^2	CFA Loadings	R^2 (CFA)
MOT1	0.70	0.50	0.50	0.66	0.44
MOT2	0.64	0.41	0.59	0.60	0.36
MOT3	0.72	0.52	0.48	0.69	0.47
ISE1	0.61	0.37	0.63	0.56	0.31
ISE2	0.42	0.18	0.82	0.38	0.14
ISE3	0.55	0.30	0.70	0.50	0.25
ISE4	0.65	0.43	0.57	0.60	0.36
GTM1	0.77	0.59	0.41	0.73	0.53
GTM2	0.71	0.51	0.49	0.67	0.44
GTM3	0.63	0.40	0.60	0.59	0.35
GTM4	0.73	0.54	0.46	0.68	0.47
COG1	0.73	0.53	0.47	0.70	0.48
COG2	0.71	0.50	0.50	0.68	0.46
COG3	0.62	0.38	0.62	0.58	0.34
COG4	0.72	0.52	0.48	0.68	0.47
COG5	0.68	0.47	0.53	0.64	0.41
COG6	0.70	0.49	0.51	0.66	0.44
SE1	0.65	0.42	0.58	0.61	0.37
SE2	0.42	0.18	0.82	0.38	0.14
SE3	0.57	0.32	0.68	0.53	0.28
SE4	0.64	0.41	0.59	0.60	0.36
SE5	0.60	0.36	0.64	0.56	0.32
SE6	0.64	0.41	0.59	0.60	0.36
MET1	0.79	0.63	0.37	0.75	0.56
MET2	0.80	0.64	0.36	0.76	0.58
MET3	0.81	0.66	0.34	0.77	0.59
MET4	0.80	0.64	0.36	0.76	0.58
MET5	0.73	0.53	0.47	0.69	0.47
MET6	0.82	0.66	0.34	0.78	0.60
MET7	0.74	0.54	0.46	0.69	0.48
MET8	0.81	0.66	0.34	0.77	0.60

Note. h^2 , or communality, represents the proportion of variance in a given observed variable that can be accounted for by the factors extracted in the EFA.

u^2 , or uniqueness, is the proportion of variance in an observed variable that is unique to it, not shared with other variables through the common factors in the model.

Supplementary Material F
Frequentist Scale Reliability Statistics

	Average interitem		
	McDonald's ω	Cronbach's α	correlation
AC-g	0.955	0.955	0.409
95% CI lower bound	0.953	0.952	0.394
95% CI upper bound	0.958	0.957	0.424
MOT	0.783	0.783	0.547
95% CI lower bound	0.767	0.768	0.524
95% CI upper bound	0.799	0.797	0.57
ISE	0.757	0.751	0.438
95% CI lower bound	0.738	0.735	0.414
95% CI upper bound	0.773	0.766	0.46
GTM			
Point estimate	0.826	0.825	0.546
95% CI lower bound	0.814	0.814	0.526
95% CI upper bound	0.838	0.836	0.566
COG	0.888	0.887	0.566
95% CI lower bound	0.881	0.88	0.549
95% CI upper bound	0.895	0.893	0.585
MET			
Point estimate	0.927	0.926	0.613
95% CI lower bound	0.922	0.922	0.596
95% CI upper bound	0.931	0.931	0.63
SE	0.843	0.841	0.47
95% CI lower bound	0.832	0.831	0.45
95% CI upper bound	0.853	0.85	0.489

Supplementary Material G*Frequentist Individual Item Reliability Statistics for Unidimensional Scales*

Item	If item dropped		Item-rest correlation
	McDonald's ω	Cronbach's α	
MOT1	0.698	0.698	0.628
MOT2	0.727	0.728	0.601
MOT3	0.693	0.693	0.634
ISE1	0.699	0.683	0.566
ISE2	0.781	0.776	0.409
ISE3	0.671	0.666	0.596
ISE4	0.647	0.644	0.636
GTM1	0.77	0.762	0.689
GTM2	0.754	0.752	0.711
GTM3	0.828	0.824	0.561
GTM4	0.782	0.779	0.649
COG1	0.859	0.858	0.755
COG2	0.856	0.855	0.775
COG3	0.881	0.88	0.622
COG4	0.865	0.862	0.73
COG5	0.88	0.879	0.621
COG6	0.869	0.866	0.704
MET1	0.917	0.917	0.746
MET2	0.916	0.916	0.761
MET3	0.915	0.915	0.772
MET4	0.916	0.916	0.766
MET5	0.92	0.92	0.713
MET6	0.914	0.913	0.792
MET7	0.922	0.922	0.692
MET8	0.917	0.917	0.753
SE1	0.816	0.81	0.644
SE2	0.844	0.842	0.482
SE3	0.828	0.823	0.575
SE4	0.795	0.793	0.725
SE5	0.819	0.816	0.609
SE6	0.803	0.801	0.683

Supplementary Material H*Frequentist Individual Item Reliability Statistics for the AC-g Scale*

Item	If item dropped		Item-rest correlation
	McDonald's ω	Cronbach's α	
MOT1	0.953	0.953	0.649
MOT2	0.954	0.954	0.587
MOT3	0.953	0.953	0.673
ISE1	0.954	0.954	0.556
ISE2	0.956	0.956	0.377
ISE3	0.955	0.954	0.499
ISE4	0.954	0.954	0.601
GTM1	0.953	0.953	0.717
GTM2	0.953	0.953	0.657
GTM3	0.954	0.954	0.57
GTM4	0.953	0.953	0.672
COG1	0.953	0.953	0.67
COG2	0.953	0.953	0.652
COG3	0.954	0.954	0.561
COG4	0.953	0.953	0.657
COG5	0.954	0.953	0.621
COG6	0.953	0.953	0.638
SE1	0.954	0.954	0.604
SE2	0.956	0.956	0.382
SE3	0.954	0.954	0.523
SE4	0.954	0.954	0.598
SE5	0.954	0.954	0.557
SE6	0.954	0.954	0.599
MET1	0.953	0.953	0.725
MET2	0.953	0.953	0.729
MET3	0.953	0.952	0.743
MET4	0.953	0.953	0.735
MET5	0.953	0.953	0.659
MET6	0.953	0.952	0.749
MET7	0.953	0.953	0.668
MET8	0.953	0.952	0.749

Supplementary Material I*Local Dependencies among Items for the Unidimensional Scales (Yen's Q3)*

	MET1	MET2	MET3	MET4	MET5	MET6	MET7	MET8
MET1	1	0.115	-0.058	-0.207	-0.15	-0.251	-0.155	-0.208
MET2	0.115	1	0.147	-0.174	-0.201	-0.329	-0.278	-0.223
MET3	-0.058	0.147	1	0.022	-0.226	-0.32	-0.315	-0.302
MET4	-0.207	-0.174	0.022	1	-0.159	-0.225	-0.201	-0.115
MET5	-0.15	-0.201	-0.226	-0.159	1	0.103	0.063	-0.119
MET6	-0.251	-0.329	-0.32	-0.225	0.103	1	0.102	-0.038
MET7	-0.155	-0.278	-0.315	-0.201	0.063	0.102	1	0.126
MET8	-0.208	-0.223	-0.302	-0.115	-0.119	-0.038	0.126	1
	COG1	COG2	COG3	COG4	COG5	COG6		
COG1	1	-0.117	-0.239	-0.342	-0.213	-0.309		
COG2	-0.117	1	-0.284	-0.363	-0.228	-0.335		
COG3	-0.239	-0.284	1	0.256	-0.215	-0.112		
COG4	-0.342	-0.363	0.256	1	-0.145	-0.091		
COG5	-0.213	-0.228	-0.215	-0.145	1	0.196		
COG6	-0.309	-0.335	-0.112	-0.091	0.196	1		
	SE1	SE2	SE3	SE4	SE5	SE6		
SE1	1	0.106	0.12	-0.385	-0.16	-0.244		
SE2	0.106	1	0.005	-0.23	-0.18	-0.212		
SE3	0.12	0.005	1	-0.285	-0.173	-0.236		
SE4	-0.385	-0.23	-0.285	1	-0.173	-0.257		
SE5	-0.16	-0.18	-0.173	-0.173	1	-0.175		
SE6	-0.244	-0.212	-0.236	-0.257	-0.175	1		
	ISE1	ISE2	ISE3	ISE4				
ISE1	1	-0.004	-0.299	-0.335				
ISE2	-0.004	1	-0.206	-0.247				
ISE3	-0.299	-0.206	1	-0.433				
ISE4	-0.335	-0.247	-0.433	1				
	GTM1	GTM2	GTM3	GTM4				
GTM1	1	-0.456	0.085	-0.361				
GTM2	-0.456	1	-0.392	-0.322				
GTM3	0.085	-0.392	1	-0.19				
GTM4	-0.361	-0.322	-0.19	1				
	MOT1	MOT2	MOT3					
MOT1	1	-0.403	-0.435					
MOT2	-0.403	1	-0.391					
MOT3	-0.435	-0.391	1					

Supplementary Material J
Local Dependencies among Items for the AC-g Scale (Yen's Q3)

	MOT1	MOT2	MOT3	ISE1	ISE2	ISE3	ISE4	GTM1	GTM2	GTM3	GTM4	COG1	COG2	COG3	COG4	COG5	COG6	SE1	SE2	SE3	SE4	SE5	SE6	MET1	MET2	MET3	MET4	MET5	MET6	MET7	MET8
MOT1	1	0.232	0.229	0.007	-0.012	-0.058	-0.002	0.009	-0.041	0.012	-0.118	-0.059	-0.075	-0.021	-0.07	-0.029	0.026	-0.04	0.016	0.014	0.029	-0.043	0.036	-0.066	-0.024	-0.004	-0.075	-0.084	-0.131	-0.096	-0.117
MOT2	0.232	1	0.225	0.15	-0.032	-0.093	-0.048	0.078	-0.061	0.103	-0.059	0.041	-0.006	0.034	0.016	-0.012	-0.01	0.005	-0.001	0.09	-0.025	0.025	-0.043	-0.129	-0.1	-0.145	-0.111	-0.034	-0.089	-0.002	-0.1
MOT3	0.229	0.225	1	0.119	-0.012	-0.039	-0.008	0.156	0.004	0.126	-0.041	-0.009	0.013	0.026	0.003	-0.051	0.042	0.024	0.032	-0.026	-0.003	-0.026	-0.008	-0.167	-0.168	-0.194	-0.199	-0.062	-0.099	-0.042	-0.064
ISE1	0.007	0.15	0.119	1	0.161	0.267	0.263	0.195	0.258	-0.115	0.092	-0.069	-0.038	-0.185	-0.197	0.014	-0.083	0.037	0.024	0.043	-0.013	-0.119	0.016	-0.108	-0.188	-0.154	-0.103	-0.131	-0.111	-0.016	-0.098
ISE2	-0.012	-0.032	-0.012	0.161	1	0.193	0.162	0.041	0.156	0.048	0.128	-0.019	-0.05	-0.108	-0.11	-0.007	-0.105	-0.066	0.131	-0.035	-0.053	-0.072	-0.005	-0.009	-0.097	-0.094	-0.112	-0.077	-0.072	0.043	-0.071
ISE3	-0.058	-0.093	-0.039	0.267	0.193	1	0.473	0.013	0.249	-0.194	0.247	-0.1	-0.116	-0.185	-0.134	0.092	-0.162	0.053	0.025	-0.044	0.129	-0.034	0.135	-0.078	-0.169	-0.165	-0.069	-0.129	-0.074	-0.019	-0.029
ISE4	-0.002	-0.048	-0.008	0.263	0.162	0.473	1	-0.037	0.196	-0.22	0.374	-0.227	-0.189	-0.268	-0.269	0.058	-0.162	0.095	0.012	-0.008	0.159	-0.004	0.181	-0.095	-0.171	-0.114	-0.066	-0.112	-0.12	0.018	0.015
GTM1	0.009	0.078	0.156	0.195	0.041	0.013	-0.037	1	0.271	0.231	0.036	-0.016	-0.072	0.016	-0.058	-0.145	-0.089	-0.015	0.026	0.032	0	-0.005	-0.028	-0.12	-0.166	-0.188	-0.165	-0.094	-0.095	-0.011	-0.112
GTM2	-0.041	-0.061	0.004	0.258	0.156	0.249	0.196	0.271	1	0.086	0.407	-0.143	-0.153	-0.206	-0.209	-0.078	-0.168	-0.086	-0.072	-0.083	-0.021	-0.135	0.029	-0.033	-0.171	-0.137	-0.067	-0.118	-0.057	0.031	-0.115
GTM3	0.012	0.103	0.126	-0.115	0.048	-0.194	-0.22	0.231	0.086	1	0.038	0.011	-0.07	0.157	0.138	-0.074	0.003	-0.192	-0.12	-0.124	-0.173	-0.029	-0.199	-0.047	-0.009	0.019	-0.092	0.038	0.001	0.081	-0.076
GTM4	-0.118	-0.059	-0.041	0.092	0.128	0.247	0.374	0.036	0.407	0.038	1	-0.235	-0.208	-0.233	-0.235	-0.053	-0.203	-0.078	-0.094	-0.109	-0.011	-0.088	0.04	0.005	-0.139	-0.115	-0.06	0.015	0.012	0.152	-0.057
COG1	-0.059	0.041	-0.009	-0.069	-0.019	-0.1	-0.227	-0.016	-0.143	0.011	-0.235	1	0.541	0.198	0.224	0.145	0.197	-0.04	-0.042	0.025	-0.081	-0.049	-0.155	-0.167	-0.093	-0.146	-0.124	-0.129	-0.09	-0.132	0.022
COG2	-0.075	-0.006	0.013	-0.038	-0.05	-0.116	-0.189	-0.072	-0.153	-0.07	-0.208	0.541	1	0.233	0.285	0.216	0.247	-0.057	-0.044	-0.008	-0.076	-0.053	-0.147	-0.153	-0.1	-0.151	-0.104	-0.11	-0.094	-0.138	0.004
COG3	-0.021	0.034	0.026	-0.185	-0.108	-0.185	-0.268	0.016	-0.206	0.157	-0.233	0.198	0.233	1	0.423	-0.011	0.135	-0.019	-0.061	-0.017	-0.076	0.037	-0.172	-0.144	0.015	0.01	-0.084	0.011	-0.045	-0.055	-0.043
COG4	-0.07	0.016	0.003	-0.197	-0.11	-0.134	-0.269	-0.058	-0.209	0.138	-0.235	0.224	0.285	0.423	1	0.093	0.223	-0.043	-0.093	-0.048	-0.179	-0.005	-0.228	-0.096	0.026	-0.04	-0.097	0.034	-0.044	-0.092	-0.065
COG5	-0.029	-0.012	-0.051	0.014	-0.007	0.092	0.058	-0.145	-0.078	-0.074	-0.053	0.145	0.216	-0.011	0.093	1	0.337	0.057	-0.054	-0.022	-0.015	-0.089	-0.052	-0.09	-0.086	-0.073	-0.072	-0.14	-0.136	-0.152	-0.045
COG6	0.026	-0.01	0.042	-0.083	-0.105	-0.162	-0.162	-0.089	-0.168	0.003	-0.203	0.197	0.247	0.135	0.223	0.337	1	-0.016	-0.059	-0.019	-0.135	-0.11	-0.18	-0.102	0.014	-0.079	-0.075	-0.067	-0.093	-0.178	-0.039
SE1	-0.04	0.005	0.024	0.037	-0.066	0.053	0.095	-0.015	-0.086	-0.192	-0.078	-0.04	-0.057	-0.019	-0.043	0.057	-0.016	1	0.332	0.33	0.246	0.191	0.215	-0.078	-0.144	-0.105	-0.083	-0.112	-0.153	-0.131	-0.065
SE2	0.016	-0.001	0.032	0.024	0.131	0.025	0.012	0.026	-0.072	-0.12	-0.094	-0.042	-0.044	-0.061	-0.093	-0.054	-0.059	0.332	1	0.218	0.219	0.138	0.193	-0.072	-0.119	-0.101	-0.051	-0.087	-0.092	-0.116	-0.037
SE3	0.014	0.09	-0.026	0.043	-0.035	-0.044	-0.008	0.032	-0.083	-0.124	-0.109	0.025	-0.008	-0.017	-0.048	-0.022	-0.019	0.33	0.218	1	0.253	0.159	0.192	-0.052	-0.036	-0.042	-0.09	-0.116	-0.159	-0.123	-0.128
SE4	0.029	-0.025	-0.003	-0.013	-0.053	0.129	0.159	0	-0.021	-0.173	-0.011	-0.081	-0.076	-0.076	-0.179	-0.015	-0.135	0.246	0.219	0.253	1	0.387	0.504	-0.022	-0.063	-0.072	-0.106	-0.164	-0.188	-0.156	-0.133
SE5	-0.043	0.025	-0.026	-0.119	-0.072	-0.034	-0.004	-0.005	-0.135	-0.029	-0.088	-0.049	-0.053	0.037	-0.005	-0.089	-0.11	0.191	0.138	0.159	0.387	1	0.352	0.025	-0.02	-0.063	-0.053	-0.054	-0.128	-0.122	-0.097
SE6	0.036	-0.043	-0.008	0.016	-0.005	0.135	0.181	-0.028	0.029	-0.199	0.04	-0.155	-0.147	-0.172	-0.228	-0.052	-0.18	0.215	0.193	0.192	0.504	0.352	1	0.072	-0.108	-0.076	-0.028	-0.145	-0.154	-0.145	-0.114
MET1	-0.066	-0.129	-0.167	-0.108	-0.009	-0.078	-0.095	-0.12	-0.033	-0.047	0.005	-0.167	-0.153	-0.144	-0.096	-0.09	-0.102	-0.078	-0.072	-0.052	-0.022	0.025	0.072	1	0.247	0.171	0.033	0.017	-0.069	-0.051	-0.125
MET2	-0.024	-0.1	-0.168	-0.188	-0.097	-0.169	-0.171	-0.166	-0.171	-0.009	-0.139	-0.093	-0.1	0.015	0.026	-0.086	0.014	-0.144	-0.119	-0.036	-0.063	-0.02	-0.108	0.247	1	0.351	0.112	-0.032	-0.087	-0.139	-0.115
MET3	-0.004	-0.145	-0.194	-0.154	-0.094	-0.165	-0.114	-0.188	-0.137	0.019	-0.115	-0.146	-0.151	0.01	-0.04	-0.073	-0.079	-0.105	-0.101	-0.042	-0.072	-0.063	-0.076	0.171	0.351	1	0.285	0.03	-0.074	-0.118	-0.145
MET4	-0.075	-0.111	-0.199	-0.103	-0.112	-0.069	-0.066	-0.165	-0.067	-0.092	-0.06	-0.124	-0.104	-0.084	-0.097	-0.072	-0.075	-0.083	-0.051	-0.09	-0.106	-0.053	-0.028	0.033	0.112	0.285	1	0.075	0.014	-0.085	0.018
MET5	-0.084	-0.034	-0.062	-0.131	-0.077	-0.129	-0.112	-0.094	-0.118	0.038	0.015	-0.129	-0.11	0.011	0.034	-0.14	-0.067	-0.112	-0.087	-0.116	-0.164	-0.054	-0.145	0.017	-0.032	0.03	0.075	1	0.284	0.2	-0.005
MET6	-0.131	-0.089	-0.099	-0.111	-0.072	-0.074	-0.12	-0.095	-0.057	0.001	0.012	-0.09	-0.094	-0.045	-0.044	-0.136	-0.093	-0.153	-0.092	-0.159	-0.188	-0.128	-0.154	-0.069	-0.087	-0.074	0.014	0.284	1	0.209	0.079
MET7	-0.096	-0.002	-0.042	-0.016	0.043	-0.019	0.018	-0.011	0.031	0.081	0.152	-0.132	-0.138	-0.055	-0.092	-0.152	-0.178	-0.131	-0.116	-0.123	-0.156	-0.122	-0.145	-0.051	-0.139	-0.118	-0.085	0.2	0.209	1	0.203
MET8	-0.117	-0.1	-0.064	-0.098	-0.071	-0.029	0.015	-0.112	-0.115	-0.076	-0.057	0.022	0.004	-0.043	-0.065	-0.045	-0.039	-0.065	-0.037	-0.128	-0.133	-0.097	-0.114	-0.125	-0.115	-0.145	0.018	-0.005	0.079	0.203	1

Note. Min: -0.269; 1st Qu.: -0.108; Median: 0.048; Mean: -0.021; 3rd Qu.: 0.022; Max: 0.541

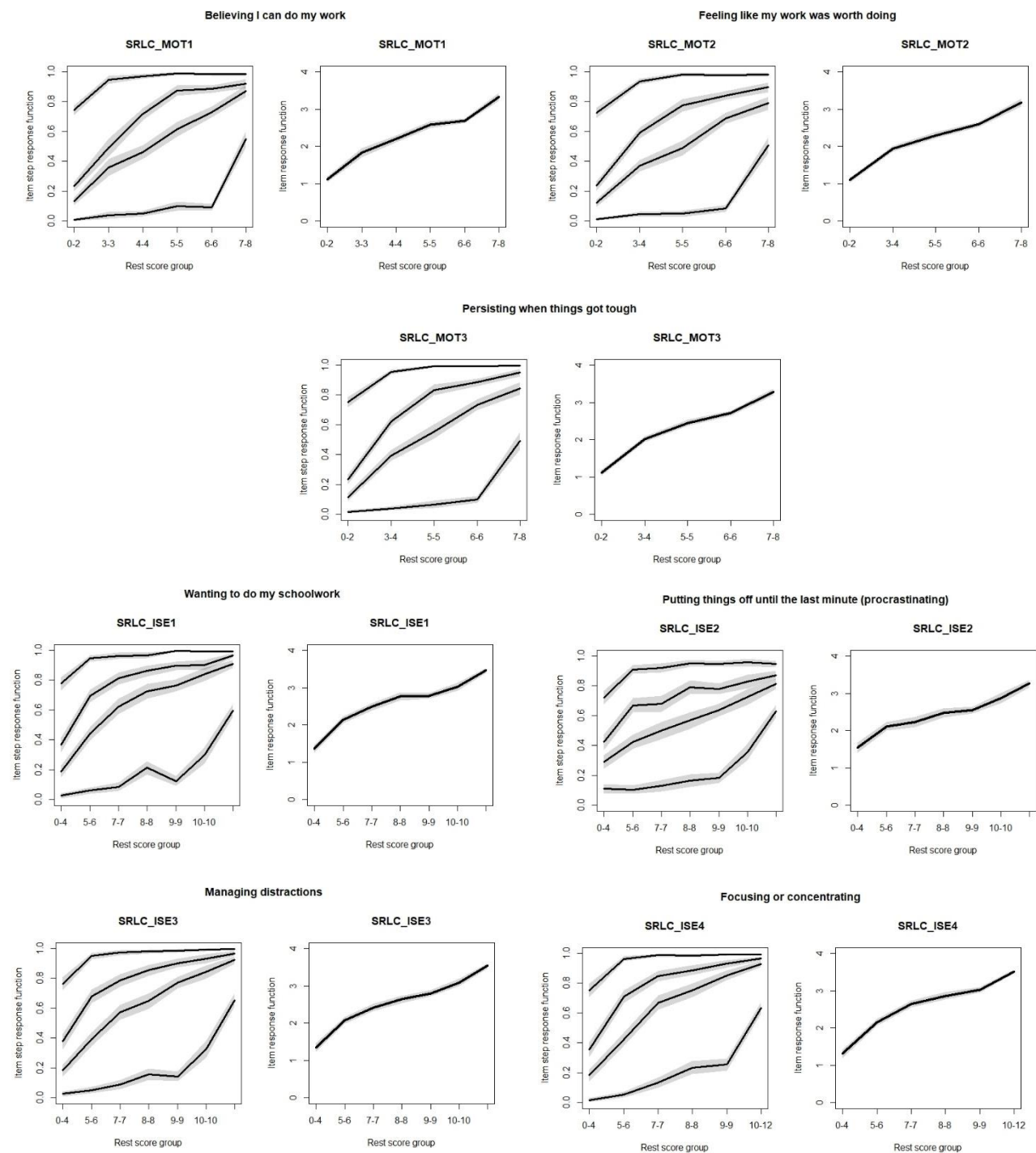
Supplementary Material K*Monotonicity Indices for the Unidimensional Subscales*

	ItemH	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	zmax	#zsig	crit
MOT1	0.56	60	0	0	0	0	0	0	0	0
MOT2	0.55	40	0	0	0	0	0	0	0	0
MOT3	0.57	40	0	0	0	0	0	0	0	0
ISE1	0.46	84	1	0.01	0.09	0.09	0.0011	3.13	1	33
ISE2	0.36	84	0	0	0	0	0	0	0	0
ISE3	0.48	84	0	0	0	0	0	0	0	0
ISE4	0.51	60	0	0	0	0	0	0	0	0
GTM1	0.61	84	0	0	0	0	0	0	0	0
GTM2	0.63	55	0	0	0	0	0	0	0	0
GTM3	0.54	60	0	0	0	0	0	0	0	0
GTM4	0.58	55	0	0	0	0	0	0	0	0
COG1	0.64	74	0	0	0	0	0	0	0	0
COG2	0.65	73	0	0	0	0	0	0	0	0
COG3	0.56	84	1	0.01	0.03	0.03	0.0004	1.04	0	0
COG4	0.63	78	0	0	0	0	0	0	0	0
COG5	0.56	84	0	0	0	0	0	0	0	0
COG6	0.6	72	0	0	0	0	0	0	0	0
SE1	0.5	84	0	0	0	0	0	0	0	0
SE2	0.39	84	1	0.01	0.04	0.04	0.0005	2.65	1	28
SE3	0.46	84	0	0	0	0	0	0	0	0
SE4	0.56	84	0	0	0	0	0	0	0	0
SE5	0.49	78	0	0	0	0	0	0	0	0
SE6	0.53	84	0	0	0	0	0	0	0	0
MET1	0.62	98	0	0	0	0	0	0	0	0
MET2	0.65	112	0	0	0	0	0	0	0	0
MET3	0.64	99	0	0	0	0	0	0	0	0
MET4	0.64	99	0	0	0	0	0	0	0	0
MET5	0.61	106	0	0	0	0	0	0	0	0
MET6	0.66	92	0	0	0	0	0	0	0	0
MET7	0.59	105	0	0	0	0	0	0	0	0
MET8	0.63	84	0	0	0	0	0	0	0	0

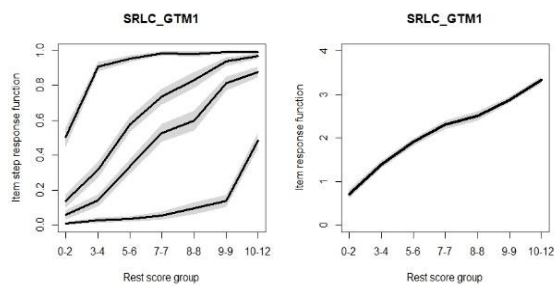
Note. itemH: item's scalability coefficient; #ac: number of active pairs that have been investigated in an item, #vi: the number of active pairs that resulted in a violation, #vi/#ac: the number of violations per active pair, maxvi: the largest violation; sum: the sum of the violations; sum/#ac: the sum of the violations per active pair; zmax: the largest value of the test statistic; #zsig: the number of statistically significant violations; and crit: an overall statistic, for which higher values indicate a worse fit. (see Crişan, et al., 2022; Koopman & Van der Ark, 2023; Molenaar & Sijtsma, 2000; Van der Ark, 2007).

Supplementary Material L

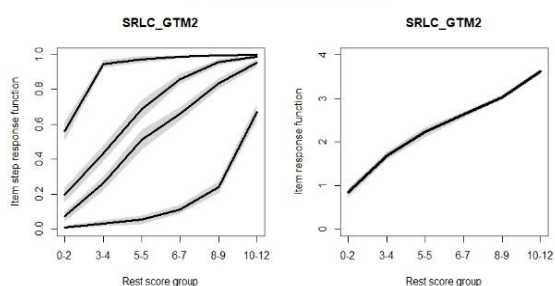
Monotonicity Plots for the Unidimensional Subscales



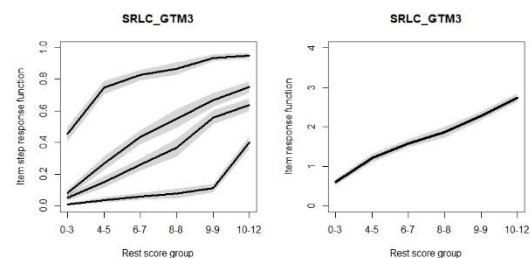
Keeping my commitments or goals



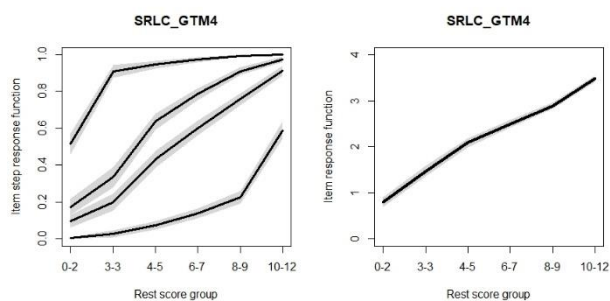
Managing my time, tasks, or goals



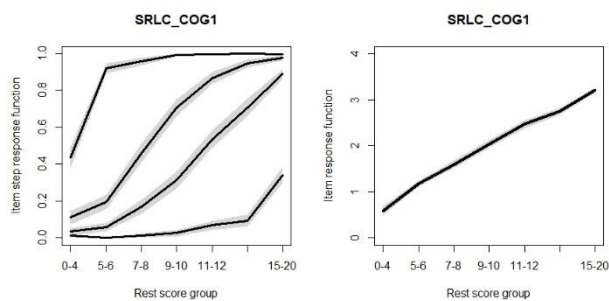
Meeting deadlines



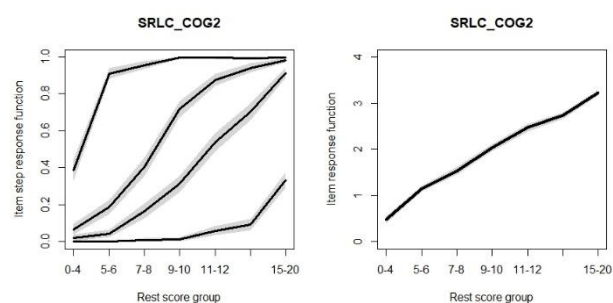
Organizing time and tasks



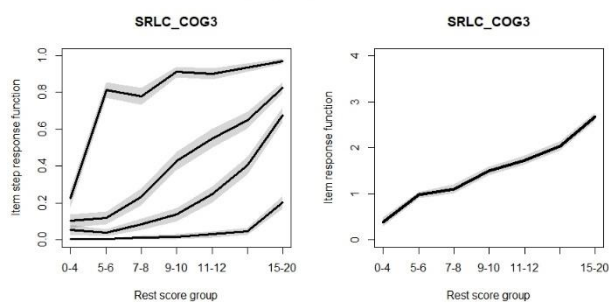
Applying my learning



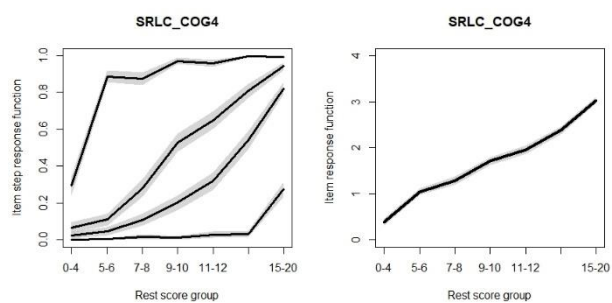
Elaborating and connecting what I learned



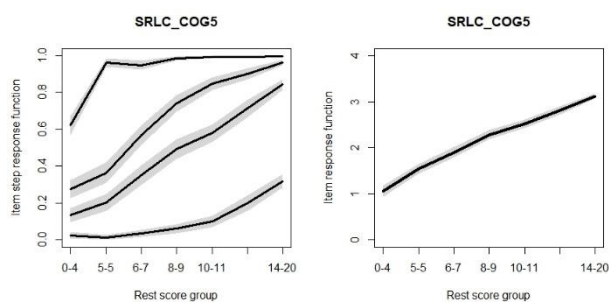
Language and communication



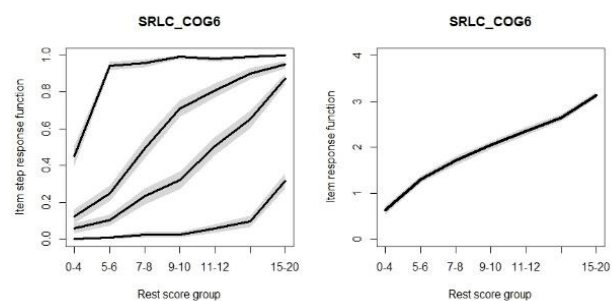
Identifying important ideas



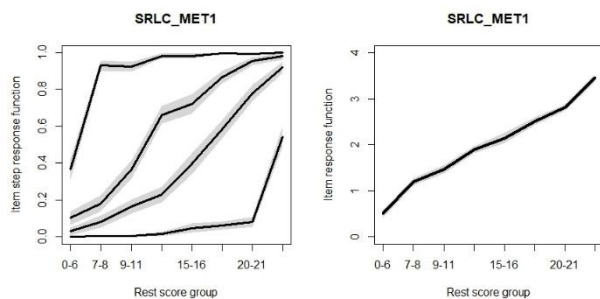
Remembering things that I learned



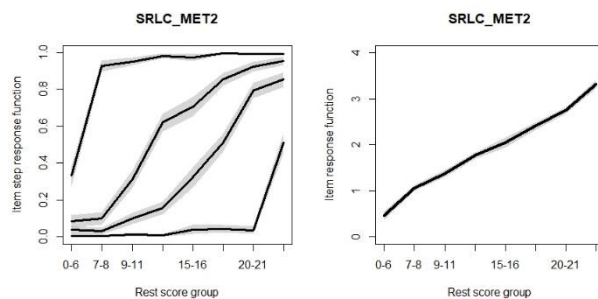
Understanding learning materials and concepts



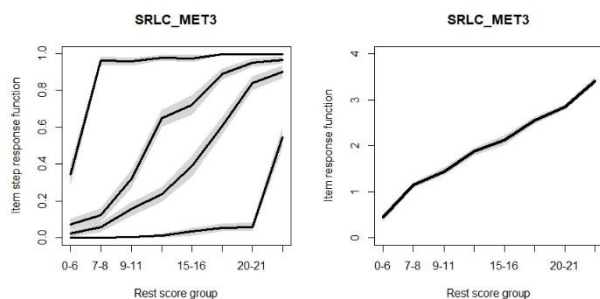
Breaking tasks down



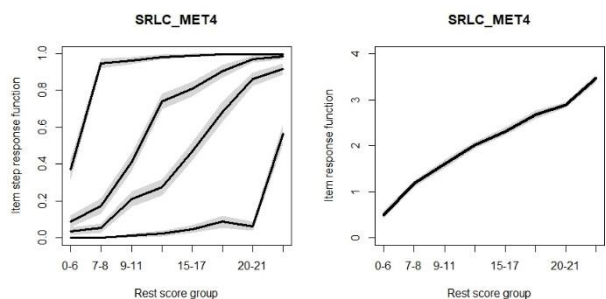
Interpreting tasks and expectations



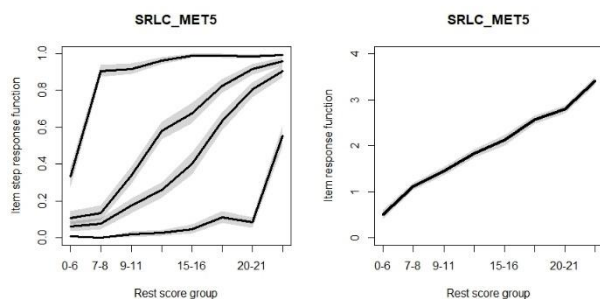
Knowing how to tackle tasks



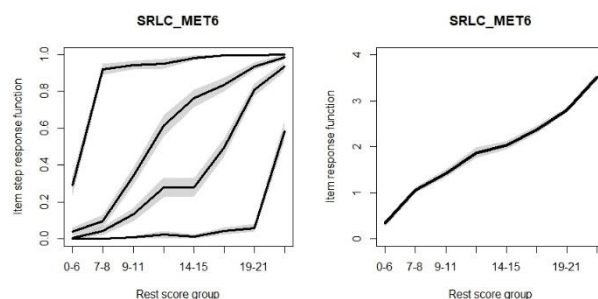
Knowing how or when to fix strategies/study skills



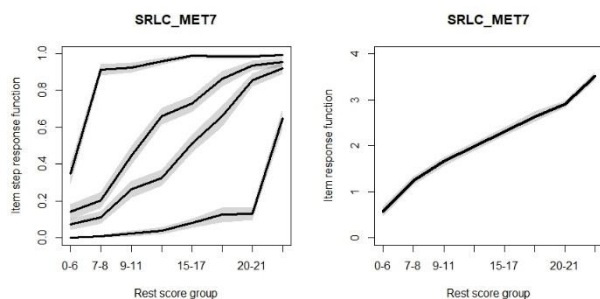
Prioritizing or figuring out what is important



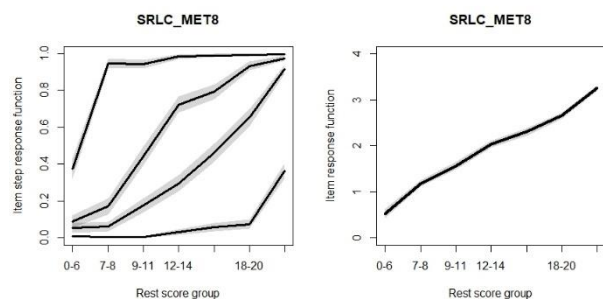
Monitoring my task or learning progress

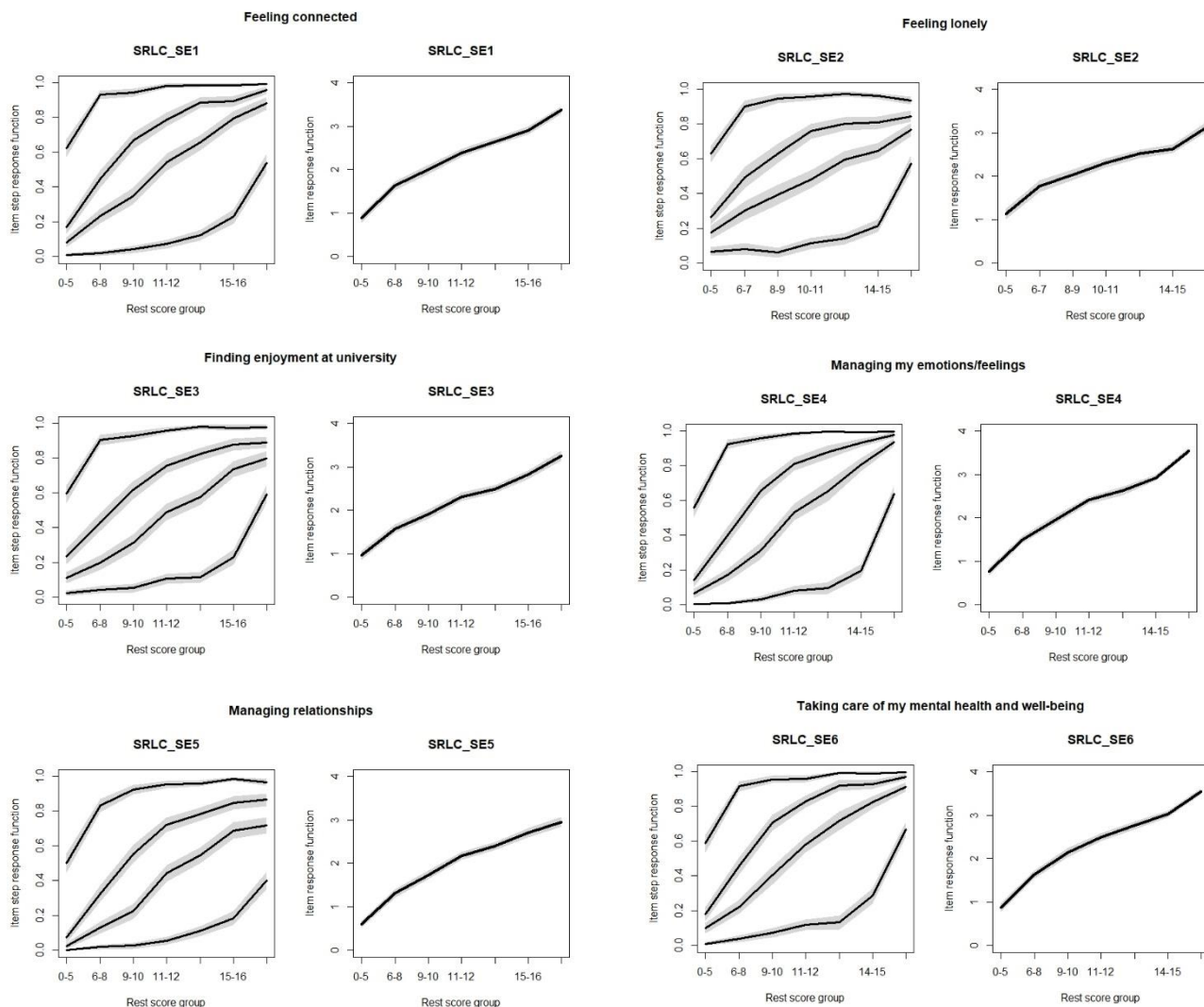


Setting goals for studying



Applying strategies/techniques properly





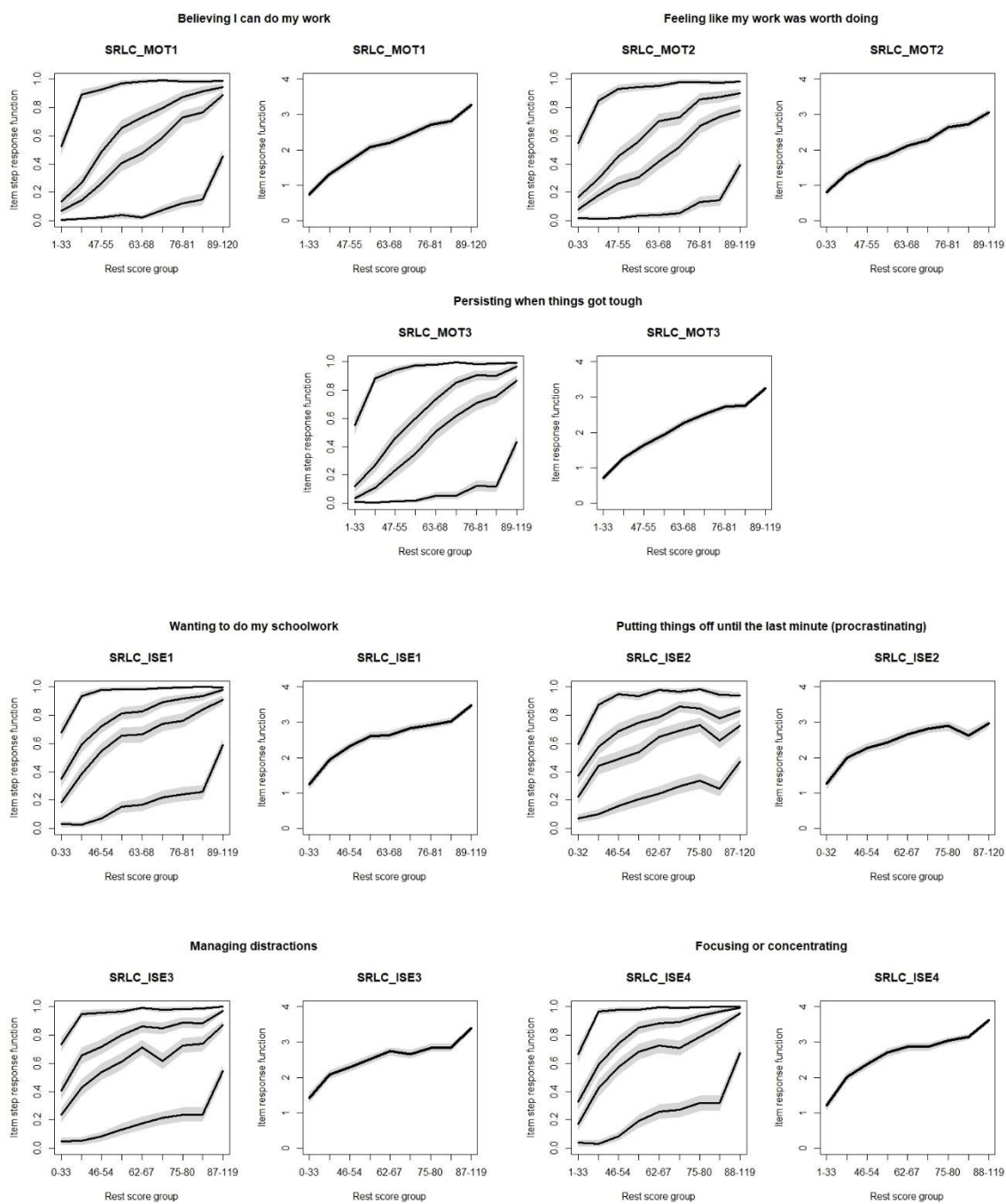
Supplementary Material M
Monotonicity Indices for the AC-g Scale

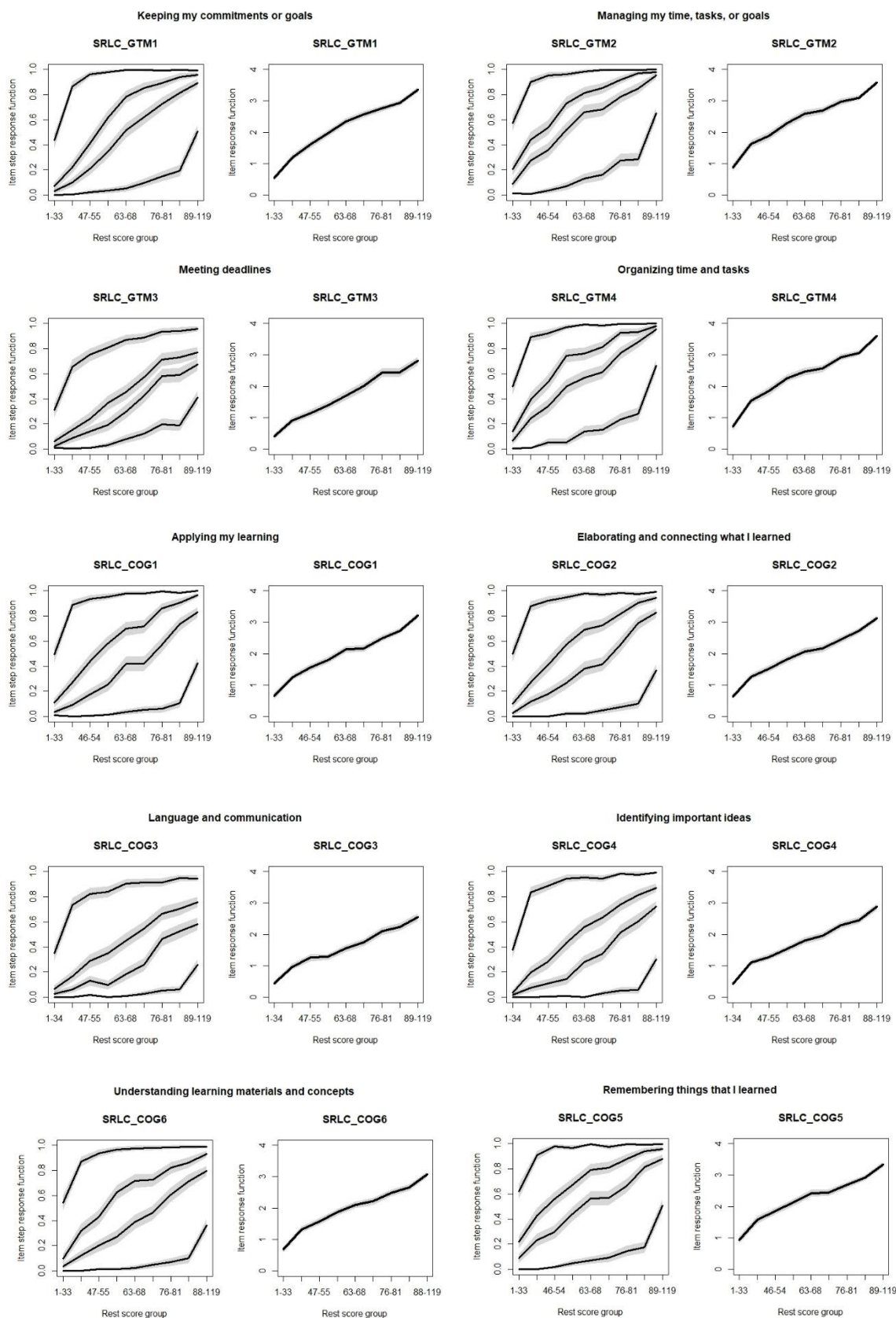
	ItemH	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	zmax	#zsig	crit
MOT1	0.45	144	0	0	0	0	0	0	0	0
MOT2	0.4	144	0	0	0	0	0	0	0	0
MOT3	0.46	144	0	0	0	0	0	0	0	0
ISE1	0.4	137	0	0	0	0	0	0	0	0
ISE2	0.27	144	10	0.07	0.11	0.56	0.0039	2.71	7	78
ISE3	0.36	144	1	0.01	0.1	0.1	0.0007	2.32	1	34
ISE4	0.44	129	0	0	0	0	0	0	0	0
GTM1	0.49	144	0	0	0	0	0	0	0	0
GTM2	0.46	144	0	0	0	0	0	0	0	0
GTM3	0.41	144	0	0	0	0	0	0	0	0
GTM4	0.47	144	0	0	0	0	0	0	0	0
COG1	0.46	129	0	0	0	0	0	0	0	0
COG2	0.45	137	0	0	0	0	0	0	0	0
COG3	0.41	136	1	0.01	0.03	0.03	0.0003	1.1	0	7
COG4	0.47	129	0	0	0	0	0	0	0	0
COG5	0.43	144	0	0	0	0	0	0	0	0
COG6	0.44	136	0	0	0	0	0	0	0	0
SE1	0.42	144	0	0	0	0	0	0	0	0
SE2	0.27	144	6	0.04	0.08	0.29	0.002	2.71	3	56
SE3	0.36	144	1	0.01	0.03	0.03	0.0002	0.93	0	8
SE4	0.41	144	0	0	0	0	0	0	0	0
SE5	0.39	144	0	0	0	0	0	0	0	0
SE6	0.42	144	1	0.01	0.07	0.07	0.0005	1.81	1	25
MET1	0.5	123	0	0	0	0	0	0	0	0
MET2	0.51	129	0	0	0	0	0	0	0	0
MET3	0.51	124	0	0	0	0	0	0	0	0
MET4	0.5	144	0	0	0	0	0	0	0	0
MET5	0.46	144	0	0	0	0	0	0	0	0
MET6	0.51	144	0	0	0	0	0	0	0	0
MET7	0.46	136	0	0	0	0	0	0	0	0
MET8	0.51	115	0	0	0	0	0	0	0	0

Note. itemH: item's scalability coefficient; #ac: number of active pairs that have been investigated in an item, #vi: the number of active pairs that resulted in a violation, #vi/#ac: the number of violations per active pair, maxvi: the largest violation; sum: the sum of the violations; sum/#ac: the sum of the violations per active pair; zmax: the largest value of the test statistic; #zsig: the number of statistically significant violations; and crit: an overall statistic, for which higher values indicate a worse fit. (see Crişan, et al., 2022; Koopman & Van der Ark, 2023; Molenaar & Sijtsma, 2000; Van der Ark, 2007).

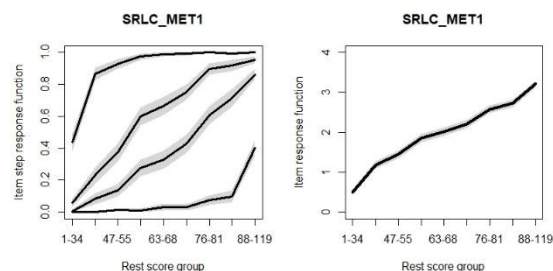
Supplementary Material N

Monotonicity Plots for the AC-g Scale

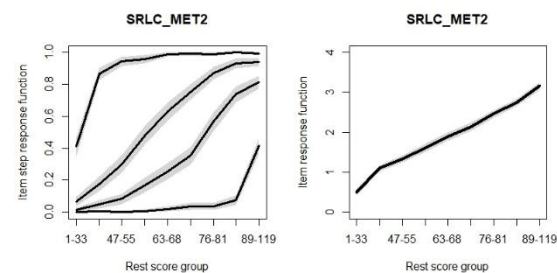




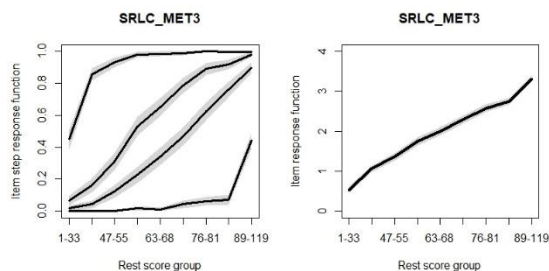
Breaking tasks down



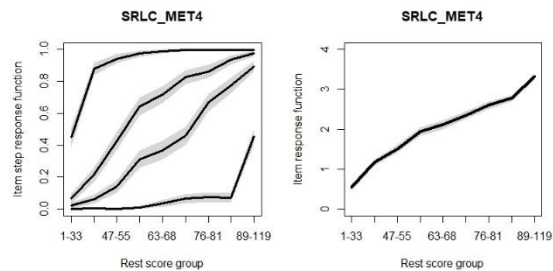
Interpreting tasks and expectations



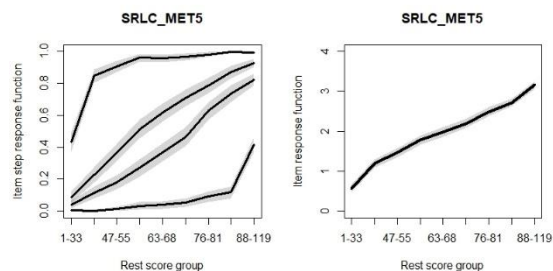
Knowing how to tackle tasks



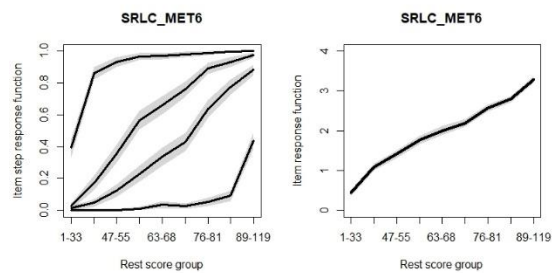
Knowing how or when to fix strategies/study skills



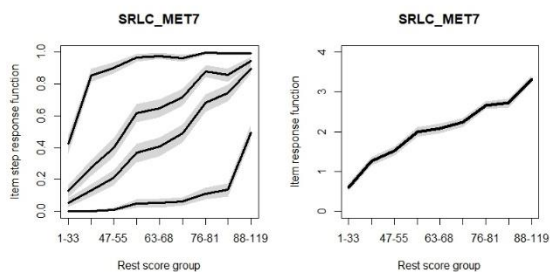
Prioritizing or figuring out what is important



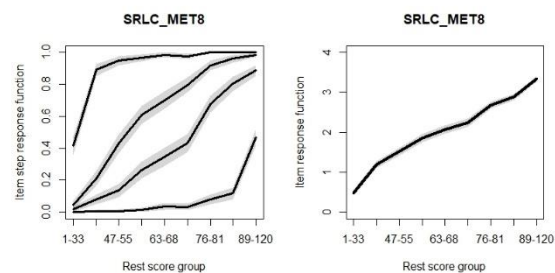
Monitoring my task or learning progress

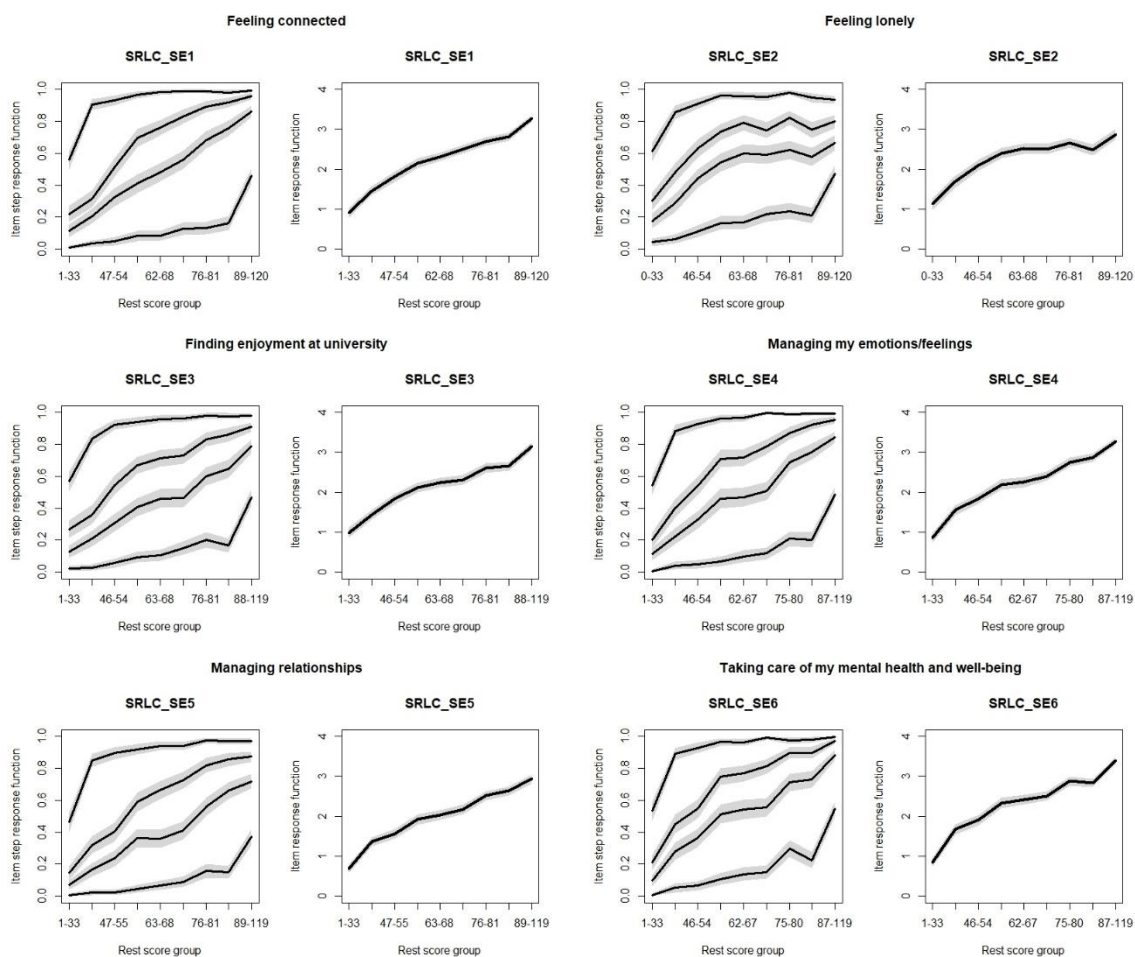


Setting goals for studying



Applying strategies/techniques properly





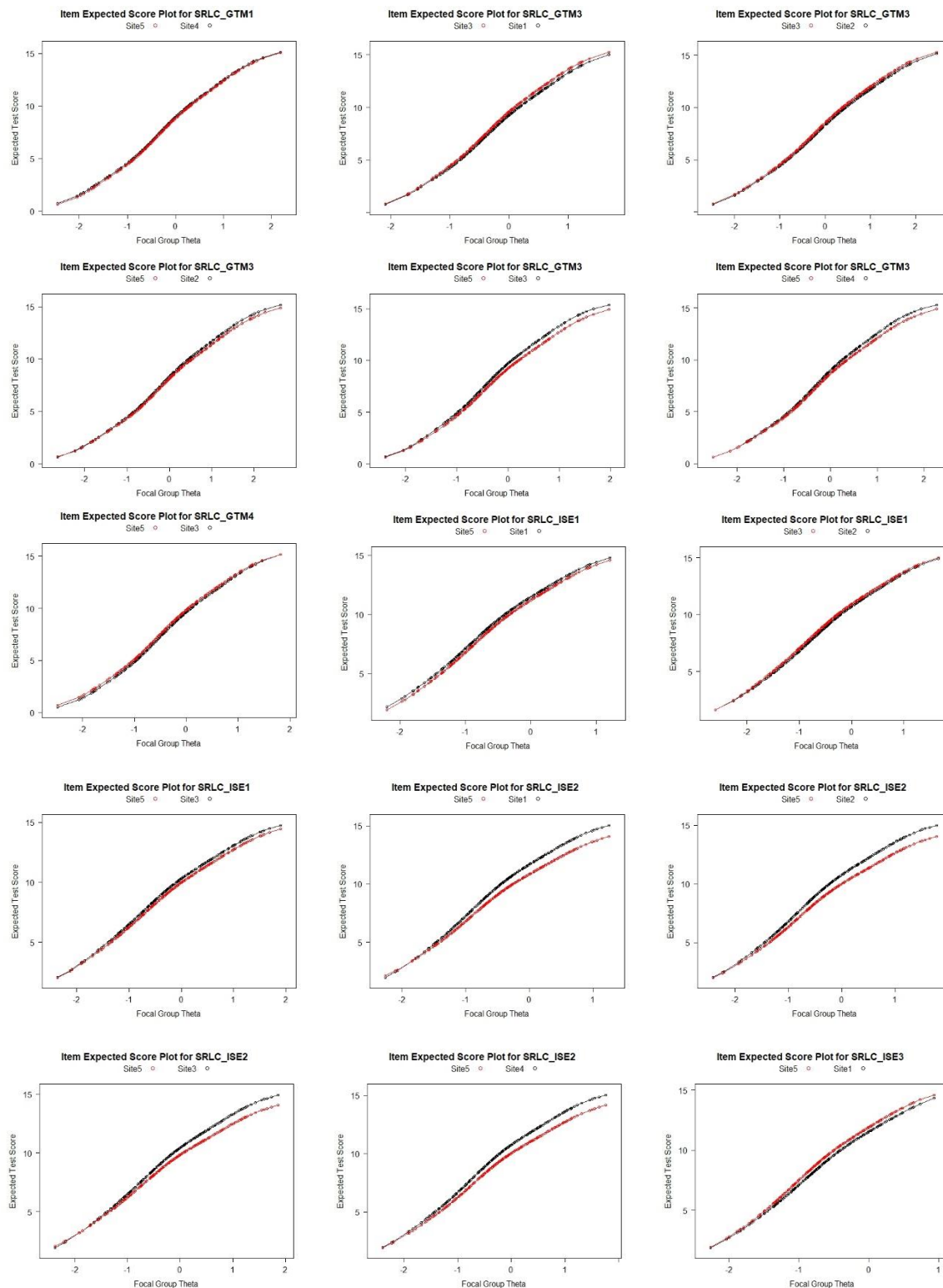
Supplementary Material O*Differential Item Functioning across the Participating Universities for the Unidimensional Scales*

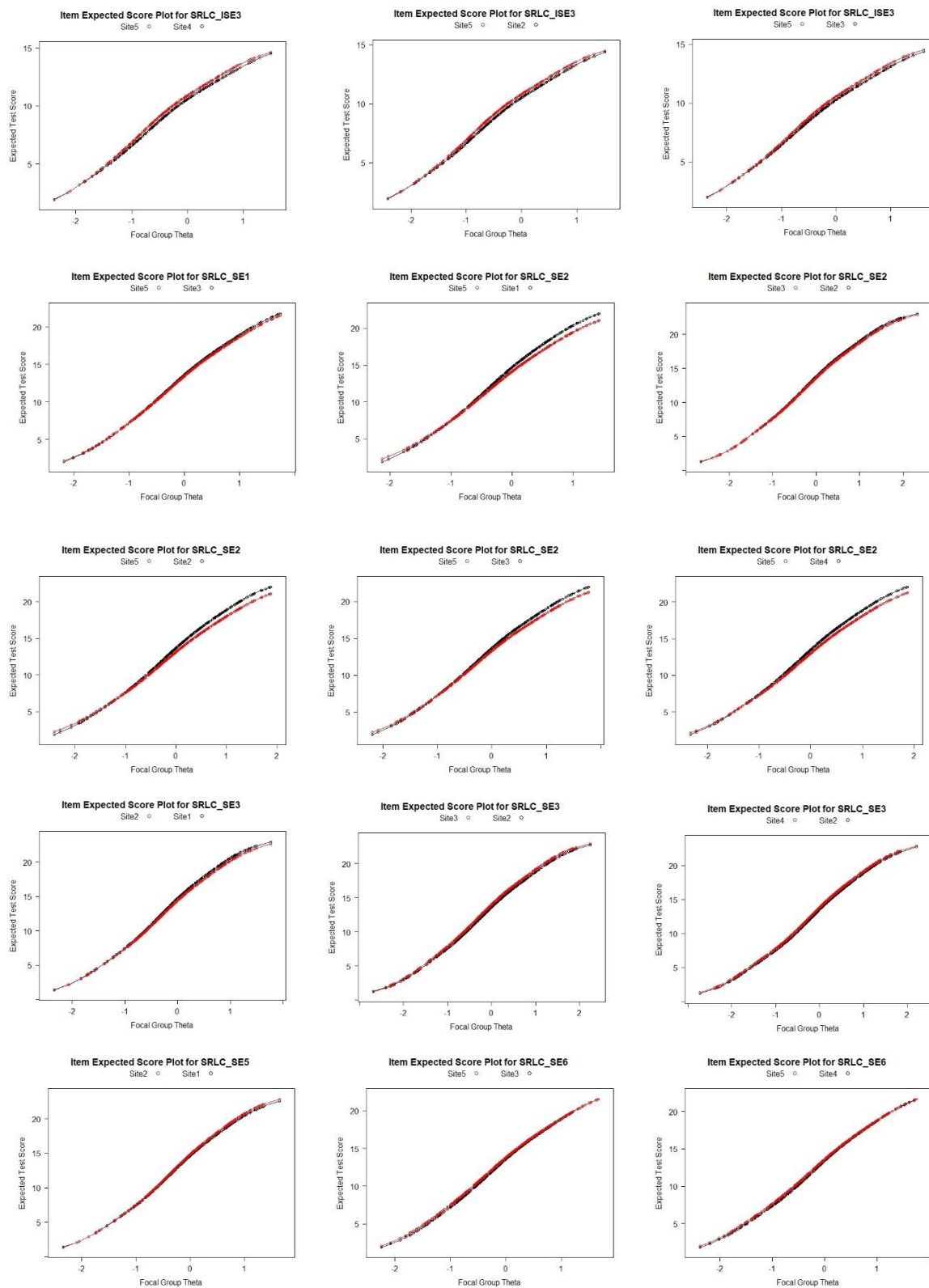
name	SIDS	UIDS	SIDN	UIDN	ESSD	theta.of.max.D	max.D	mean.ES.foc	mean.ES.ref
GTM3_13	0.281	0.281	0.266	0.266	0.344	0.601	0.362	2.033	1.752
GTM3_23	0.203	0.203	0.206	0.206	0.242	0.822	0.248	2.034	1.831
GTM3_25	-0.212	0.213	-0.216	0.216	-0.262	1.596	-0.326	1.504	1.716
GTM3_35	-0.421	0.421	-0.419	0.419	-0.5	0.972	-0.589	1.503	1.924
GTM4_35	0.204	0.204	0.206	0.206	0.208	-1.669	0.254	2.396	2.192
GTM3_45	-0.303	0.304	-0.305	0.305	-0.364	1.403	-0.484	1.504	1.807
COG3_13	0.228	0.237	0.216	0.228	0.271	0.827	0.378	1.856	1.629
COG4_15	0.192	0.197	0.182	0.189	0.246	-0.23	0.253	1.625	1.434
COG4_25	0.192	0.243	0.205	0.251	0.229	-1.754	0.385	1.625	1.433
COG6_25	-0.198	0.198	-0.199	0.199	-0.242	-0.194	-0.248	1.7	1.899
COG5_35	0.231	0.231	0.22	0.22	0.319	-0.615	0.291	2.181	1.95
SE3_12	-0.258	0.259	-0.245	0.247	-0.336	0.914	-0.401	1.971	2.228
SE5_12	0.162	0.163	0.16	0.161	0.22	1.316	0.307	2.089	1.927
SE2_15	-0.455	0.484	-0.424	0.499	-0.819	1.276	-0.945	1.906	2.361
SE2_23	-0.162	0.165	-0.16	0.161	-0.216	1.245	-0.262	2.241	2.403
SE3_23	0.329	0.329	0.327	0.327	0.434	0.12	0.387	2.267	1.938
SE3_24	0.273	0.273	0.271	0.271	0.372	0.083	0.31	2.204	1.931
SE2_25	-0.503	0.529	-0.486	0.526	-0.908	1.705	-0.97	1.906	2.409
SE2_35	-0.339	0.362	-0.325	0.371	-0.675	1.777	-0.726	1.907	2.245
SE6_35	0.213	0.214	0.197	0.2	0.249	-0.671	0.314	2.439	2.226
SE2_45	-0.472	0.482	-0.455	0.473	-0.936	1.803	-0.81	1.906	2.378

Note. The notation "GTM3_13" refers to the difference in responses to item GTM3 between Site 1 and Site 3.

Supplementary Material P

Differential Item Functioning Across the Participating Universities for the Unidimensional Subscales





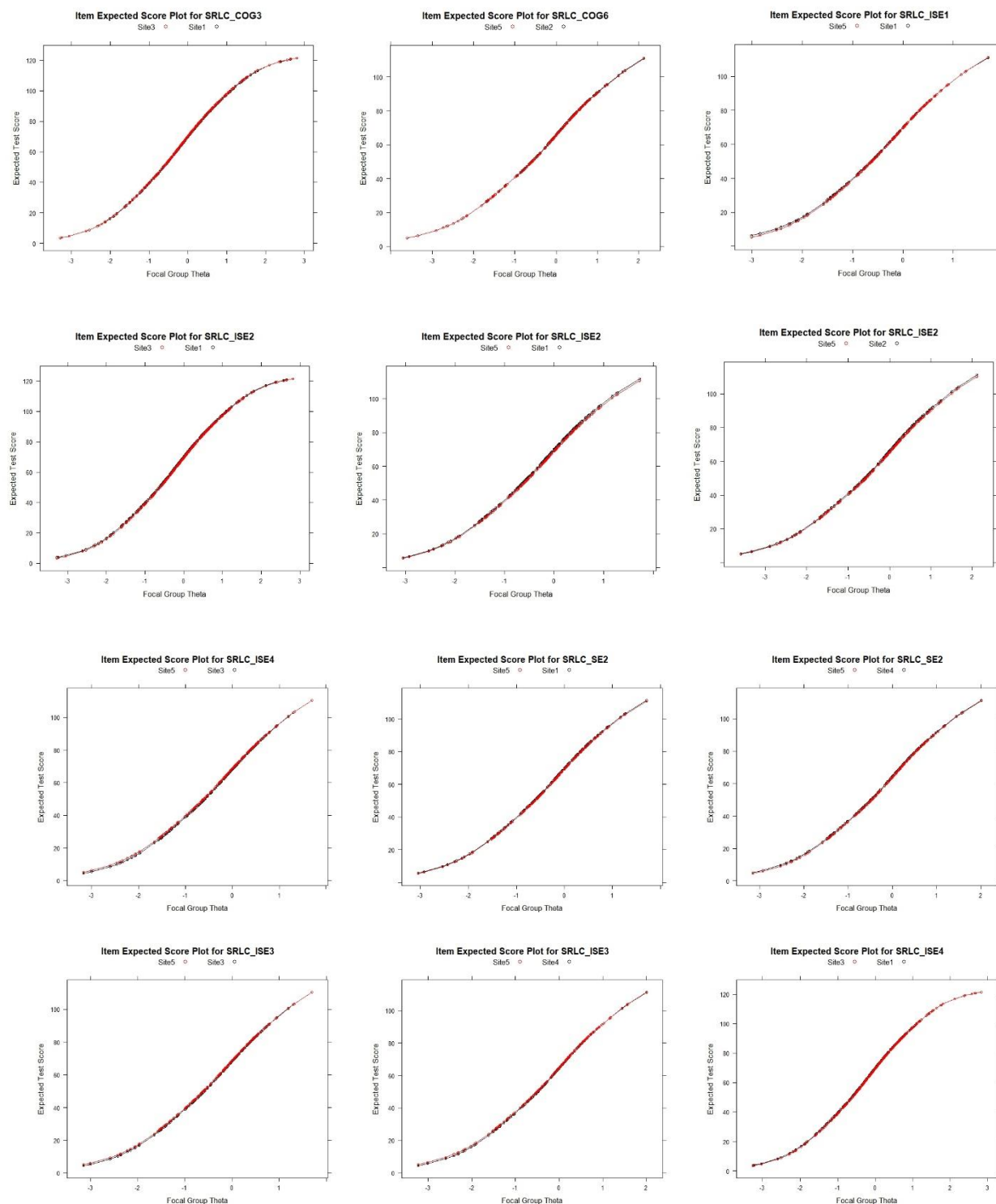
Supplementary Material Q*Differential Item Functioning across the Participating Universities for the AC-g Scale*

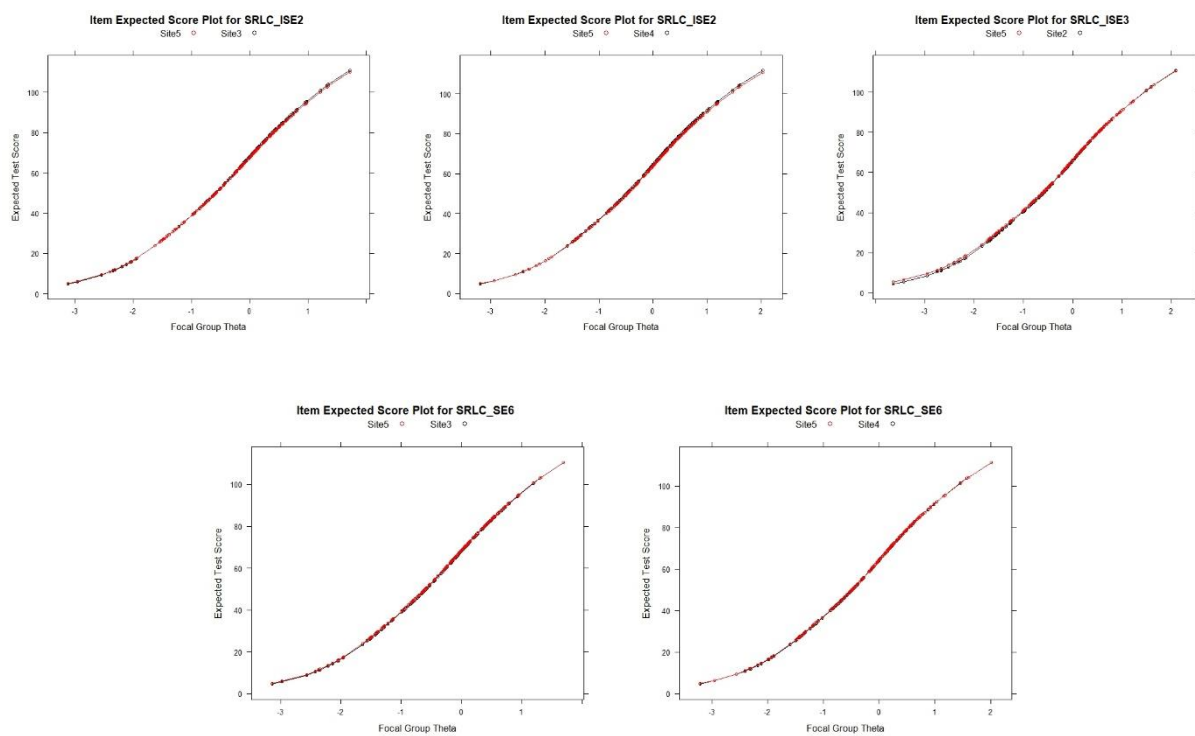
	SIDS	UIDS	SIDN	UIDN	ESSD	theta.of.max.D	max.D	mean.ES.foc	mean.ES.ref
ISE1_15	-0.293	0.294	-0.302	0.305	-0.462	-2.107	-0.68	2.387	2.68
ISE2_13	-0.369	0.369	-0.364	0.364	-0.644	-0.438	-0.412	2.402	2.771
ISE2_15	-0.706	0.706	-0.685	0.686	-1.467	0.852	-0.839	1.922	2.628
ISE2_35	-0.32	0.325	-0.315	0.319	-0.676	1.809	-0.546	1.922	2.242
ISE2_25	-0.614	0.614	-0.613	0.613	-1.386	1.125	-0.692	1.922	2.535
ISE2_45	-0.541	0.541	-0.539	0.539	-1.216	1.171	-0.621	1.922	2.463
ISE3_25	0.295	0.306	0.299	0.311	0.542	-2.918	0.761	2.743	2.448
ISE3_35	0.45	0.45	0.45	0.45	0.884	-2.511	0.753	2.743	2.293
ISE3_45	0.358	0.361	0.361	0.364	0.679	-2.573	0.699	2.743	2.384
ISE4_13	-0.3	0.3	-0.292	0.292	-0.403	-0.646	-0.371	2.548	2.848
ISE4_35	0.407	0.414	0.408	0.418	0.656	-2.425	0.833	2.73	2.323
COG3_13	0.298	0.3	0.285	0.285	0.396	0.912	0.431	1.861	1.563
COG6_25	-0.31	0.31	-0.307	0.307	-0.412	1.882	-0.437	1.715	2.025
SE2_15	-0.443	0.448	-0.432	0.438	-0.954	1.657	-0.749	1.908	2.351
SE2_45	-0.367	0.367	-0.365	0.365	-0.915	1.909	-0.485	1.907	2.274
SE6_35	0.382	0.383	0.378	0.381	0.573	-1.719	0.627	2.426	2.045
SE6_45	0.292	0.301	0.295	0.306	0.43	-1.967	0.553	2.426	2.134

Note. The notation "ISE1_15" refers to the difference in responses to item ISE1 between Site 1 and Site 5.

Supplementary Material R

Differential Item Functioning Across the Participating Universities for the AC-g Scale





Supplementary Material S*Item Fit Measures for the Unidimensional Subscales*

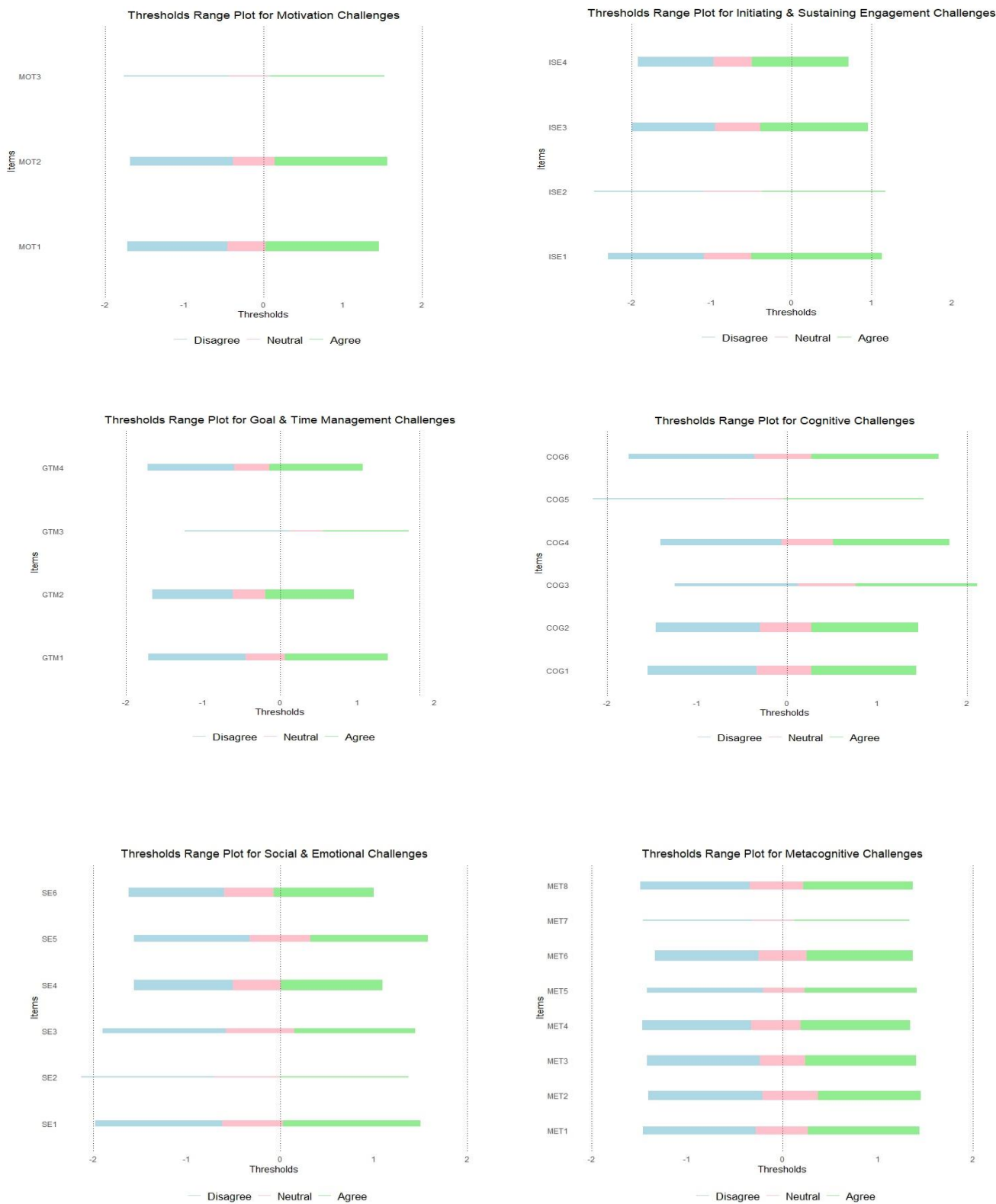
	Item	S_X ²	df	RMSEA	p
MOT1	Believing I can do my work	70.22	11.00	0.05	0.0000
MOT2	Feeling like my work was worth doing	36.68	12.00	0.03	0.0000
MOT3.	Persisting when things got tough	25.77	11.00	0.02	0.0070
ISE1.	Wanting to do my schoolwork	45.73	27.00	0.02	0.0140
ISE2.	Putting things off until the last minute (procrastinating)	133.89	30.00	0.04	0.0000
ISE3	Managing distractions	33.70	25.00	0.01	0.1150
ISE4	Focusing or concentrating	49.98	25.00	0.02	0.0020
GTM1	Keeping my commitments or goals	43.59	26.00	0.02	0.0170
GTM2	Managing my time, tasks, or goals	66.05	22.00	0.03	0.0000
GTM3	Meeting deadlines	44.79	29.00	0.01	0.0310
GTM4	Organizing time and tasks	65.47	26.00	0.02	0.0000
COG1	Applying my learning	196.62	39.00	0.04	0.0000
COG2	Elaborating and connecting what I learned	86.89	35.00	0.02	0.0000
COG3	Language and communication	211.23	50.00	0.04	0.0000
COG4	Identifying important ideas	135.00	45.00	0.03	0.0000
COG5	Remembering things that I learned	202.44	51.00	0.03	0.0000
COG6	Understanding learning materials and concepts	139.68	46.00	0.03	0.0000
SE1	Feeling connected	136.84	53.00	0.02	0.0000
SE2	Feeling lonely	289.56	59.00	0.04	0.0000
SE3	Finding enjoyment at university	225.76	55.00	0.03	0.0000
SE4	Managing my emotions/feelings	121.93	42.00	0.03	0.0000
SE5	Managing relationships	214.55	52.00	0.03	0.0000
SE6	Taking care of my mental health and well-being	153.77	47.00	0.03	0.0000
MET1	Breaking tasks down	163.29	59.00	0.03	0.0000
MET2	Interpreting tasks and expectations	196.85	57.00	0.03	0.0000
MET3	Knowing how to tackle tasks	150.79	54.00	0.03	0.0000
MET4	Knowing how or when to fix strategies/study skills	187.03	56.00	0.03	0.0000
MET5	Prioritizing or figuring out what is important	221.55	66.00	0.03	0.0000
MET6	Monitoring my task or learning progress	152.61	56.00	0.03	0.0000
MET7	Setting goals for studying	293.42	72.00	0.03	0.0000
MET8	Applying strategies/techniques properly	157.37	58.00	0.03	0.0000

Supplementary Material T*Item Fit Measures for the AC-g Scale*

item		S_X ²	df	RMSEA	p
MOT1	Believing I can do my work	278.621	252	0.006	0.246
MOT2	Feeling like my work was worth doing	292.867	278	0.005	0.351
MOT3	Persisting when things got tough	243.22	243	0.001	0.536
ISE1	Wanting to do my schoolwork	305.99	293	0.004	0.362
ISE2	Putting things off until the last minute (procrastinating)	430.584	349	0.009	0.009
ISE3	Managing distractions	320.941	302	0.005	0.351
ISE4	Focusing or concentrating	362.524	273	0.011	0.002
GTM1	Keeping my commitments or goals	202.95	228	0	0.882
GTM2	Managing my time, tasks, or goals	256.526	260	0	0.587
GTM3	Meeting deadlines	302.979	287	0.005	0.351
GTM4	Organizing time and tasks	273.551	248	0.006	0.246
COG1	Applying my learning	259.847	237	0.006	0.254
COG2	Elaborating and connecting what I learned	229.056	243	0	0.755
COG3	Language and communication	343.053	280	0.009	0.021
COG4	Identifying important ideas	279.058	236	0.008	0.074
COG5	Remembering things that I learned	355.839	260	0.012	0.002
COG6	Understanding learning materials and concepts	299.062	246	0.009	0.036
SE1	Feeling connected	365.145	277	0.011	0.002
SE2	Feeling lonely	458.14	352	0.011	0.002
SE3	Finding enjoyment at university	354.967	310	0.007	0.095
SE4	Managing my emotions/feelings	340.888	288	0.008	0.049
SE5	Managing relationships	310.364	295	0.004	0.351
SE6	Taking care of my mental health and well-being	323.058	287	0.007	0.156
MET1	Breaking tasks down	214.97	205	0.004	0.362
MET2	Interpreting tasks and expectations	260.239	193	0.012	0.006
MET3	Knowing how to tackle tasks	201.064	189	0.005	0.351
MET4	Knowing how or when to fix strategies/study skills	258.56	197	0.011	0.009
MET5	Prioritizing or figuring out what is important	238.773	229	0.004	0.362
MET6	Monitoring my task or learning progress	201.903	193	0.004	0.362
MET7	Setting goals for studying	303.19	238	0.01	0.01
MET8	Applying strategies/techniques properly	216.516	195	0.006	0.253

Supplementary Material U

Thresholds Range Plots for the Unidimensional Scales



Note. The horizontal bars represent threshold ranges for each item. The thickness of these bars signifies their level of discrimination, with thicker bars indicating greater discrimination. The leftward floating position of the bars suggests a high rate of student agreement with the items, while a relative position to the right indicates a high rate of disagreement with the items. Vertical dotted lines at $\theta = -2, 0$, and 2 are drawn to facilitate comparisons among the items and to illustrate the latent ranges where the items fail to measure reliably.

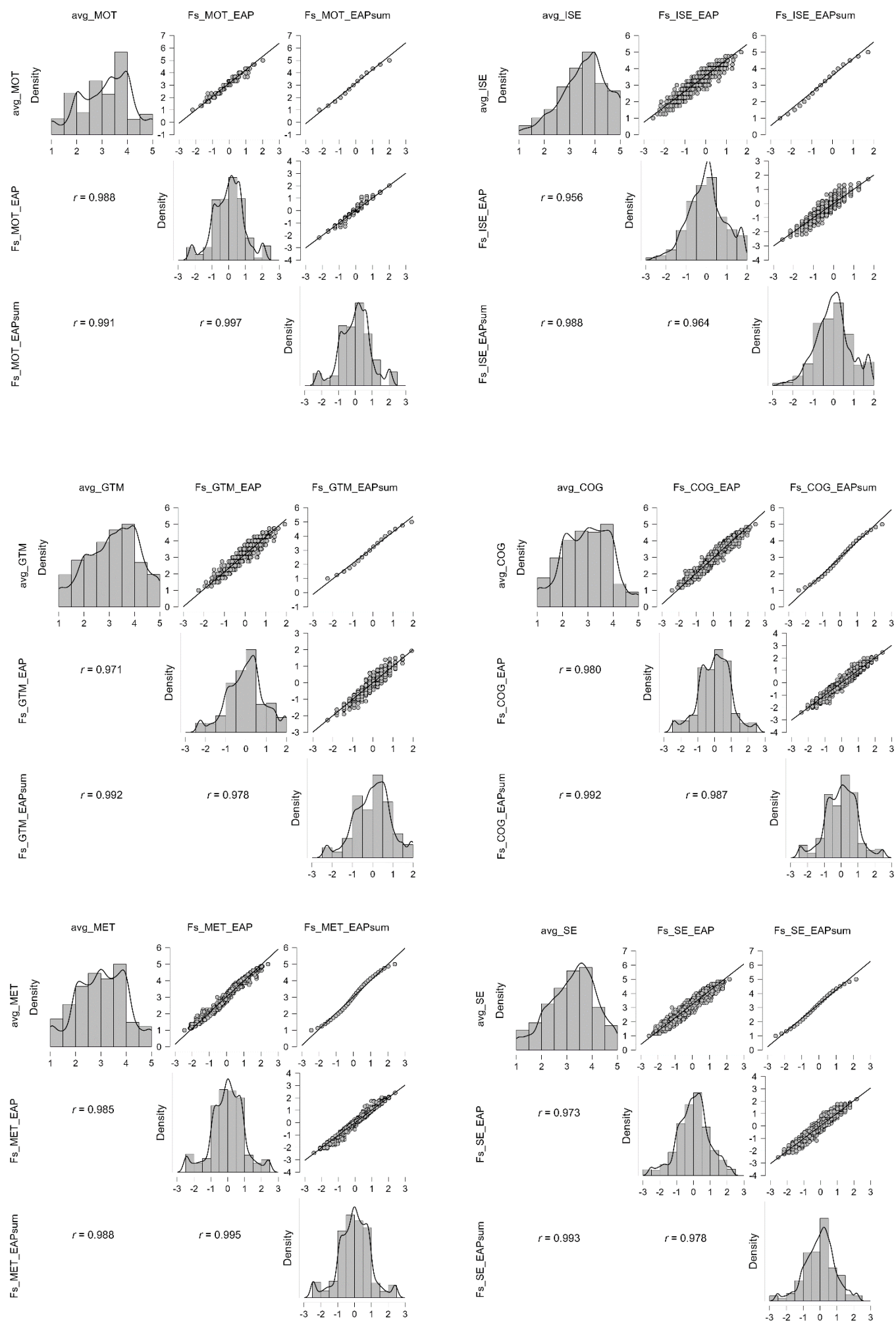
Supplementary Material V*Stratified alpha for the AC-g Scale*

	scale	Items	alpha	mean.tot	var.tot	alpha.stratified
1	AC-g	31	0.955	98.09	574.119	0.965
2	COG	6	0.887	17.869	30.124	
3	GTM	4	0.825	12.744	16.364	
4	ISE	4	0.751	14.31	12.757	
5	MET	8	0.926	24.25	57.132	
6	MOT	3	0.783	9.457	8.612	
7	SE	6	0.841	19.436	30.16	

Note. See Rae (2007).

Note. Abbreviation: MOT: Motivation; ISE: Initiating & Sustaining; GTM: Goal & Time Management; COG: Cognitive; MET: Metacognitive; SE: Social & Emotional.

Supplementary Material W
Distributions and Linear associations between the Composite Scores



Note. Abbreviations: avg: average; EAP: Expected A Posteriori; EAPsum: Expected A Posteriori for summed scores; REG: Regression Factor Scores.

Supplementary Material X
Correlation Matrices for the Scales across Various Scoring Methods

Variable	MOT _{TF}	ISE _{TF}	GTM _{TF}	COG _{TF}	SE _{TF}	MET _{TF}	
MOT _{TF}	—						
ISE _{TF}	0.687	—					
GTM _{TF}	0.816	0.851	—				
COG _{TF}	0.756	0.539	0.683	—			
SE _{TF}	0.67	0.664	0.65	0.609	—		
MET _{TF}	0.782	0.663	0.835	0.794	0.681	—	
	MOT _{Reg}	ISE _{Reg}	GTM _{Reg}	COG _{Reg}	SE _{Reg}	MET _{Reg}	AC-g _{Reg}
MOT _{Reg}	—						
ISE _{Reg}	0.517	—					
GTM _{Reg}	0.646	0.659	—				
COG _{Reg}	0.637	0.429	0.567	—			
SE _{Reg}	0.546	0.55	0.548	0.512	—		
MET _{Reg}	0.673	0.557	0.729	0.729	0.601	—	
AC-g _{Reg}	0.792	0.676	0.827	0.829	0.728	0.944	—
	MOT _{avg}	ISE _{avg}	GTM _{avg}	COG _{avg}	SE _{avg}	MET _{avg}	AC-g _{avg}
MOT _{avg}	—						
ISE _{avg}	0.522	—					
GTM _{avg}	0.656	0.638	—				
COG _{avg}	0.641	0.432	0.594	—			
SE _{avg}	0.547	0.516	0.532	0.53	—		
MET _{avg}	0.668	0.552	0.747	0.734	0.594	—	
AC-g _{avg}	0.795	0.713	0.839	0.826	0.772	0.911	—
	MOT _{EAP}	ISE _{EAP}	GTM _{EAP}	COG _{EAP}	SE _{EAP}	MET _{EAP}	AC-g _{EAP}
MOT _{EAP}	—						
ISE _{EAP}	0.515	—					
GTM _{EAP}	0.636	0.672	—				
COG _{EAP}	0.634	0.436	0.561	—			
SE _{EAP}	0.545	0.55	0.547	0.515	—		
MET _{EAP}	0.667	0.561	0.724	0.729	0.606	—	
AC-g _{EAP}	0.779	0.672	0.814	0.827	0.723	0.947	—
	MOT _{EAPsum}	ISE _{EAPsum}	GTM _{EAPsum}	COG _{EAPsum}	SE _{EAPsum}	MET _{EAPsum}	AC-
MOT _{EAPsum}	—						
ISE _{EAPsum}	0.522	—					
GTM _{EAPsum}	0.658	0.64	—				
COG _{EAPsum}	0.642	0.438	0.604	—			
SE _{EAPsum}	0.544	0.516	0.535	0.537	—		
MET _{EAPsum}	0.664	0.559	0.745	0.74	0.602	—	
AC-g _{EAPsum}	0.79	0.71	0.836	0.834	0.772	0.909	—

Note. All estimates significant at p<.001
Abbreviations: TF: True Factor; Reg: Regression Factor Scores; EAP: Expected A Posteriori Factor Scores; EAPsum: Expected A Posteriori for Summed Scores.

Supplementary Material Y*Associations between Criterion Constructs and the Scales*

Variable 1	Variable 2	True Factors		Average		Regression FS		EAP FS.		EAPsum FS.	
		Correlation	<i>p</i>	Correlation	<i>p</i>	Correlation	<i>p</i>	Correlation	<i>p</i>	Correlation	<i>p</i>
MOT	GPA	-0.143	***	-0.142	***	-0.143	***	-0.136	***	-0.135	***
ISE	GPA	-0.089	***	-0.105	***	-0.065	0.011	-0.077	0.003	-0.114	***
GTM	GPA	-0.191	***	-0.206	***	-0.186	***	-0.178	***	-0.2	***
COG	GPA	-0.114	***	-0.119	***	-0.119	***	-0.116	***	-0.117	***
SE	GPA	-0.021	0.126	-0.021	0.42	-0.017	0.512	-0.023	0.372	-0.027	0.299
MET	GPA	-0.144	***	-0.158	***	-0.156	***	-0.145	***	-0.147	***
AC-g	GPA	-0.137	***	-0.15	***	-0.153	***	-0.147	***	-0.146	***
MOT	EGPA	-0.167	***	-0.158	***	-0.159	***	-0.15	***	-0.151	***
ISE	EGPA	-0.121	***	-0.131	***	-0.098	***	-0.1	***	-0.135	***
GTM	EGPA	-0.15	***	-0.149	***	-0.14	***	-0.132	***	-0.146	***
COG	EGPA	-0.169	***	-0.168	***	-0.171	***	-0.162	***	-0.163	***
SE	EGPA	-0.086	***	-0.089	***	-0.075	***	-0.078	***	-0.093	***
MET	EGPA	-0.162	***	-0.165	***	-0.166	***	-0.158	***	-0.157	***
AC-g	EGPA	-0.168	***	-0.176	***	-0.175	***	-0.166	***	-0.171	***
MOT	GPA_SE	-0.306	***	-0.295	***	-0.297	***	-0.286	***	-0.283	***
ISE	GPA_SE	-0.173	***	-0.177	***	-0.164	***	-0.166	***	-0.176	***
GTM	GPA_SE	-0.273	***	-0.272	***	-0.265	***	-0.257	***	-0.267	***
COG	GPA_SE	-0.258	***	-0.267	***	-0.265	***	-0.262	***	-0.267	***
SE	GPA_SE	-0.241	***	-0.248	***	-0.236	***	-0.237	***	-0.252	***
MET	GPA_SE	-0.314	***	-0.326	***	-0.329	***	-0.321	***	-0.316	***
AC-g	GPA_SE	-0.31	***	-0.33	***	-0.334	***	-0.328	***	-0.328	***

Note. Abbreviations: EGPA: GPA Expectations; GPA_SE: GPA Self-Efficacy

***: Significant at the 0.001 level ($p \leq 0.001$)

Appendix C: Study 3

Unraveling the Multidimensional Structure of the Academic Challenges Scale: Exploring First-order, Second-order, and Bifactor Models within ICM-CFA and ESEM Frameworks

Ramin Rostampour

Department of Educational Psychology and Leadership Studies

University of Victoria

This research was supported by a Social Sciences and Humanities Research Council (SSHRC) of Canada multisite partnership grant (the UniVenture consortium; PI Sherry Stewart, 895-2019-1021).

Abstract

This study addresses two major gaps in the validating constructs in the Perceived Academic Challenges Scale by exploring its multidimensionality with a sample of 2,673 university students from five Canadian universities. Prior research has not examined (a) the potential an overarching factor represents academic challenges and (b) has suppressed potential associations of items with non-target factors, possibly complicating interpretations of constructs the measure was designed to represent. To address these gaps, I (a) investigated the existence of an overarching factor using bifactor and second-order measurement models, and (b) estimates cross-loadings within the measurement model using Exploratory Structural Equation Modeling (ESEM). Additionally, to understand the scale's predictive power, I examined its associations with three criterion variables—students' GPA expectations, self-efficacy for GPA, and actual end-of-semester GPA. The results showed a general factor underlies responses. Estimated cross-loadings enhanced interpretability of the construct. A modeling approach that combined the general factor and cross-loadings (bifactor ESEM) demonstrated strongest associations with criterion variables. The reliability analysis indicated the meaning of the general factor is predominantly associated with the metacognitive challenges scale. Consequently, modifications were made to the bifactor ESEM model based on this finding. To conclude, the *Academic Challenges* scale represents constructs as intended and is reliable for assessing students' academic challenges. Suggestions are offered for further improvements to the scale.

Keywords: Academic challenges, Bifactor Measurement model, Exploratory Structural Equation Modeling, S-1 Bifactor ESEM

Background

In the field of psychology, a multidimensional construct is often modelled as a set of correlated factors. For instance, the *Academic Challenges* scale has been modelled as six correlated factors (Hadwin et al., 2024). Correlated CFA is perhaps the most common measurement model applied to operationalize multidimensional constructs.

A correlated measurement models components of a complex construct rather than the construct itself. To elaborate using a famous example, a correlated CFA for the Big 5 personality model, represents the dimensions of personality (neuroticism, extraversion, etc.), not personality. To operationalize personality, the measurement model should inform how these dimensions combine to represent personality.

In a correlated CFA model, covariances among factors are estimated as part of the measurement model, but they are not operationalized as elements that can be used in the structural model to test hypotheses. For instance, in a 6-factor model of the academic challenges, associations among the factors are operationalized as 15 bivariate correlations. However, this type of measurement model cannot identify what these covariations collectively represent and whether they can reflect a general factor reflecting the construct itself.

From a statistical point of view, it is possible to add a general factor to any model that consists of several correlated first-order factors. However, computational feasibility, a necessary condition for modeling a general factor, does not reflect conceptual and theoretical considerations essential to support whether and how the covariation among items and/or factors can be modeled as a general factor.

It is important to note that failing to model a potential general factor could lead to inflated estimates of correlations between first-order factors (Asparouhov & Muthén, 2009) as well as

regression coefficients. That causes multicollinearity among factors when attempting to use them as concurrent predictors of learning outcomes.

Two major approaches within the domain of common factor models, known as the higher-order and bifactor models, are often employed to investigate how well an overarching latent construct supersedes models involving the covariation among observations and factors (Brunner et al., 2012; Rindskopf & Rose, 1988). Both models incorporate a set of specific factors and often one general factor, but they differ in how they structure the relationships among these factors and the observed variables. The next subsection describes these two modeling approaches.

Modeling the General Factor: Second-order vs. Bifactor Measurement Models

In a second-order factor model (Figure 1, model b), one or several observed variables (items) are modelled as indicators of specific factors, and these specific factors, in turn, serve as indicators of the higher-order (general) factor (i.e., H). In this context, the specific factors represent the shared variance among observed indicators, while the higher-order factor captures the common variance among the specific factors (Rindskopf & Rose, 1988).

In a higher order measurement model, the influence of the general factor on the items is mediated through the first-order factors. This indirect influence is distributed equally across all items exogenous to a first-order factor, a phenomenon known as proportionality constraint (Reise, 2012). Also, part of the covariance among indicators of a first-order factor is absorbed by both the higher-order and first-order factor making them psychometrically redundant (Chen et al., 2006; Reise, 2012). This means that the concurrent effect of higher-order and first-order factors on any criterion outcome cannot be estimated due to psychometric overlap and multicollinearity of these factors (see Chapter 2 of the dissertation, section titled “Higher-order

Measurement Model and Bifactor Measurement Model” for a detailed description of these limitations.) However, these constraints may not always apply validly in real-world data as items within a factor can vary in how strongly they are influenced by the higher-order factor versus the specific factors, and the first-order and higher-order factors can coexist as psychometrically related yet discrete factors.

To relax these limitations, competing model, the bifactor model (Figure 1, model c) is proposed which provides a possibly more realistic representation of the data. In a bifactor model, the proportionality constraint is relaxed, allowing each item to have unique loadings on the general factor and its specific factor.

The variance of the general factor in a bifactor model is disaggregated from that of the orthogonal specific factors (Brunner et al., 2012). In a bifactor model, the variance of each observed indicator is directly explained by both the specific and general factors along with unexplained sources, often termed measurement error. In other words, the associations between items and general and specific factors are direct, making bifactor models capable of estimating co-existing and disaggregated global and specific constructs without imposing the restrictive proportionality constraints inherent in higher-order factor models. As the specific factors are orthogonal to each other in a bifactor model, the general factor represents what is shared among all observed indicators, while each specific factor represents what is shared among particular items constituting a specific factor, not already explained by the general factor (Reise, 2012).

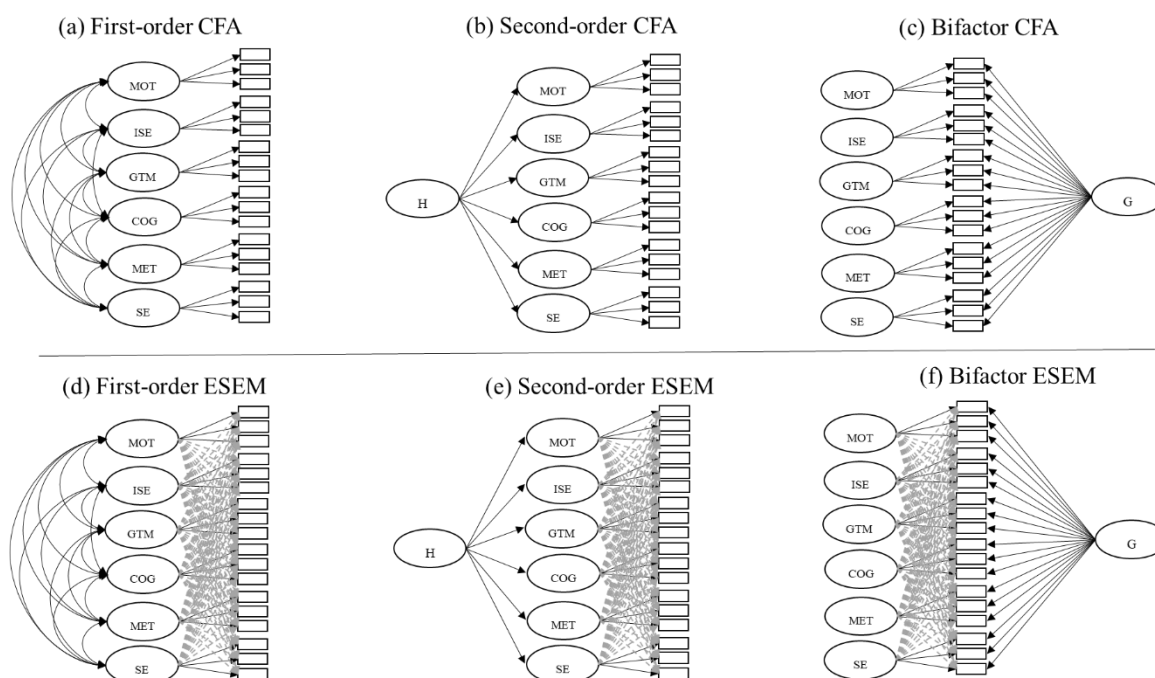
Disaggregating item variances into the variances of general and specific factors (Brunner et al., 2012; Morin et al., 2020) in a manner free from multicollinearity and conceptual redundancy allows for an examination of the incremental predictive power of coexisting specific

and general factors. This feature of bifactor models allows for testing hypotheses that demand modeling of general and specific factors as coexisting and separate factors.

Overall, bifactor modeling offers a suitable alternative for representing a hierarchically organized factor structure (Chen et al., 2006; Gillet et al., 2020; Morin et al., 2016; Morin et al., 2020; Reise, 2012). As a result, Morin (2023) and Morin et al. (2022) suggest avoiding higher-order models unless there are strong theoretical and/or empirical reasons to incorporate strict implicit proportionality constraints into the model.

Figure 1

Correlated, Second order, and Bifactor CFA and ESEM Models for the Academic Challenges Scale



Note. Unidimensional model is not included.

Abbreviations. H: Higher order; G: General; MOT: Motivation, ISE: Initiating & Sustaining Engagement; GTM: Goal & Time Management; COG: Cognitive, MET: Metacognitive Challenges; SE: Social & Emotional Challenges

What Does Exploratory Structural Equation Modeling Offer?

Beyond failing to model a potential general factor that underlies the responses, there is another construct-relevant psychometric characteristic that correlated CFA models do not take into account. Correlated CFA operates under the assumption of local independence of clusters, which is why they are also referred to as Independent Cluster Model CFA (ICM-CFA; Xiao et al., 2019). The ICM assumption posits that there is only one construct-relevant source of true variance for each observation. In simpler terms, each item exclusively loads on its target factor, and the associations between the item and all non-target factors are constrained to be zero (Xiao et al., 2019). The association between an item and non-target factors is known as cross-loading. For instance, in an independent cluster CFA (i.e., often labelled as correlated CFA), an item from the motivation challenge subscale is assumed to be exclusively caused by the motivation challenge latent factor and the associations between the item and other factors (e.g., metacognitive challenges) are forced to be zero.

The assumption of zero cross-loadings in ICM-CFA can be problematic with multidimensional instruments (Morin, 2023; Morin et al., 2016, Morin et al., 2020). Assuming that items on a subscale are not influenced by non-target factors is often unlikely and ecologically unrepresentative (Marsh et al., 2014; Morin et al., 2016; Morin et al., 2020), especially when an instrument measures conceptually related constructs such as academic challenges in which items are likely to relate to factors other than the primary ones to which they load.

Forcing the cross-loadings to zero in correlated CFA can result in inflated first-order factor correlations or inflated item-factor loadings for the general factor (Murray & Johnson, 2013). This is because the unmodelled shared variance between items and non-target factors is

likely to be pushed into factor correlations, leading to distorted estimation of relationships among the factors and among the items and the general factors. These biases in the measurement model can possibly bias the structural path estimates where the association of the measurement model and other variables is investigated (Asparouhov & Muthén, 2009; Marsh et al., 2014; Morin et al., 2020; Zhang et al., 2023).

Similarly, statistical simulation studies have demonstrated that neglecting to model even minor cross-loadings leads to biased estimates of factor correlations and regression coefficients. Conversely, permitting the free estimation of negligible cross-loadings continues to yield unbiased parameter estimates (Mai et al., 2018; Wei et al., 2022). This means that ESEM tends to produce less biased parameter estimates than ICM-CFA.

Additionally, the restrictive nature of ICM-CFA often results in models with suboptimal levels of fit, though they may represent replicable factor structures (e.g., Marsh et al., 2009; Marsh et al., 2013). Consequently, researchers frequently attempt to enhance model fit through post hoc, data-driven procedures guided by indices of misspecification, such as modification indices. A common practice involves adding covariances between item residuals, a modification that is often not theoretically grounded and therefore would lead to measurement structures that are not replicable. Correlating measurement errors typically indicates the presence of an overlooked latent factor, or it may arise from item sequencing or similar phrasing. When researchers cannot clearly identify why two items are correlated beyond their correlation caused by their target factor, the modifications made to the measurement model may merely serve to artificially improve fit indices rather than genuinely enhance the measurement model's stability and replicability.

To address these weaknesses of the ICM-CFA models, Exploratory Structural Equation Modeling (ESEM) has been proposed as an alternative analytical method (Figure 1, models *d*, *e*, *f*). ESEM allows items to primarily load on their target factor, while also permitting item loadings on non-target factors (i.e., cross-loadings). While each item primarily loads on its target factor, contributions from other non-target factors in determining the variance of the item are also allowed (Asparouhov & Muthén, 2009).

ESEM is described as a synthesis of EFA and CFA (Marsh et al., 2014), designed to overcome the limitations of the CFA approach. EFA, or Exploratory Factor Analysis, applies data-driven statistical techniques to identify the minimal number of relevant dimensions required to explain the covariation among a set of observations. In EFA, cross-loadings are freely estimated. Conversely, CFA is a theory-driven approach aimed at testing the alignment between a theoretical model and the data, where cross-loadings are constrained to zero. While EFA does not allow for the specification of a theoretical model and investigation of the global fit of the model, since factors are derived in a data-driven way and labeled post hoc, CFA's restrictive assumption regarding the cross-loadings makes it vulnerable to inflated factor correlations and deflated goodness of fit threatening the psychometric validity of the measurement model.

To conclude, ESEM combines the benefits of exploratory and confirmatory factor analysis, enabling the assessment of the global fit of a specified theoretical model while allowing for estimation of cross-loadings. These features potentially empower ESEM to provide a more realistic and ecologically valid representation of the interconnected multidimensional constructs, as it accommodates the complex relationships that often exist among items and factors in real-world data (Asparouhov & Muthén, 2009; Marsh et al., 2014). A summary of the benefits and limitations of these measurement approaches are provided in Table 1.

Table 1***Advantages and Disadvantages of CFA versus ESEM Approaches***

<i>Measurement Models</i>	<i>Pros</i>	<i>Cons</i>
CFA (Figure 1, models <i>a, b, c</i>)	Parsimoniousness	Forcing cross-loadings to zero ¹
ESEM (Figure 1, models <i>d, e, f</i>)	◦ Allows for specification of theoretical models ◦ Allows for modeling of cross loadings	◦ Less parsimony ◦ Convergence Problems

Advantages and Disadvantages of Second-order versus Bifactor Measurement Models

Second-order (Figure 1, models <i>b & e</i>)	Parsimoniousness	◦ the ratio of H factor variance to specific factor variance is the same for all indicators within a specific factor² ◦ First order and second order factors are psychometrically redundant
Bifactor (Figure 1, models <i>c & f</i>)	◦ Disaggregated g and s factors ◦ Allowing for unequal contribution of g to the items	◦ Less parsimony ◦ Vanishing specific factors due to the general factor explaining a large proportion of the variance

Note.

¹Restrictive assumption of independence of clusters

²Proportionality Constraints

Academic Challenges

Navigating academic challenges is an integral aspect of the educational journey for students. These challenges come in various forms, ranging from motivational barriers to cognitive and metacognitive difficulties, emotional hurdles, and social pressures. Understanding and addressing these challenges are important for several reasons.

First, they are part of students' experiences which can directly impact students' learning. Such an impact can be adverse or positive depending on how the challenges are addressed.

Second, by acknowledging these challenges, educators and institutions can strive to provide tailored support to enhance student learning and well-being. Understanding and addressing academic challenges has potential to enhance individual student experiences and foster a culture of inclusivity, where students feel understood and supported in their educational pursuits.

Measuring academic challenges offers a pivotal avenue for advancing research in understanding students' educational experiences. Despite extensive research on self-regulatory behaviors and perceptions linked to student success, there remains a noticeable gap in understanding the specific challenges that can potentially impede student success. By measuring these challenges, researchers increase opportunities to effectively address this disparity, enriching our comprehension of the complexities students face as they learn.

This study represents a continuation of prior validation endeavors of the Academic Challenges Scale. The Academic Challenges Scale is rooted in the Self-Regulated Learning (SRL) framework. SRL, as conceptualized across various models, emphasizes the intentional regulation of learning processes through metacognition, active engagement, and strategic actions such as planning, monitoring, enacting, and evaluating (Dunlosky & Ariel, 2011; Hadwin et al., 2017; Hadwin & Winne, 2012; Kim et al., 2020; Schunk & Greene, 2018; Winne, 1997; 2005; Zimmerman, 2013). The strategic regulation in SRL encompass a wide range of domains including cognitive, metacognitive, behavioral, motivational, and affective domains (Panadero, 2017).

Zimmerman posits that SRL is not an innate ability or skill but a transformative process that translates personal capacities into performance skills (Zimmerman, 2015). This perspective encourages learners to not only focus on outcomes but also on acquiring self-knowledge and understanding the intricacies of the ways one learns (Cambridge et al., 2013; Cleary &

Zimmerman, 2004; Pino-Pasternak et al., 2014; Whittlesey & Steiner, 2021; Zimmerman et al., 2011). Winne (1997, 2006, 2010a, 2017, 2022) conceptualizes SRL as formation of a personalized learning theory wherein learners develop hypotheses about their own learning, actively experiment with these hypotheses, reflect on the outcomes, and adjust study habits as needed.

SRL, as a goal-driven model of regulation, is commonly understood as resting on two overarching metacognitive skills: (a) monitoring, which involves identifying discrepancies between current and desired states, and (b) control, encompassing strategic actions aimed at mitigating these discrepancies (Dunlosky & Ariel, 2011; Schunk & Greene, 2018; Winne, 1997, 2005, 2013, 2017a; Winne & Hadwin, 1998; Zimmerman, 2013). The perception of a challenge is created when students become aware of a discrepancy between a desired state and the current state. These perceptions may play important roles in the process of monitoring and evaluating. While metacognitive knowledge and metacognitive awareness alone can not suffice for effective metacognitive control, they serve as fundamental pillars upon which intentional control processes are built.

In the context of the current study, academic challenges refer to perceived discrepancies between the current and desired states that learners face in managing their learning that are situated by the theoretical construct within key regulatory domains including cognitive, metacognitive, behavioral, motivational, volitional, social and emotional areas.

From a self-regulatory perspective, encountering challenges is not merely a setback but a critical step toward refining learning strategies. Challenges serve as triggers for adaptive behavior, prompting students to re-evaluate ineffective approaches and engage in more effective self-regulation methods (Hadwin et al., 2019; Järvelä et al., in press). However, prolonged, and

intense challenges can transform from growth opportunities to barriers hindering learning and engagement, underscoring the necessity of timely identification and intervention (Boekaerts, 2011; Tuominen-Soini et al., 2012; Cleary & Zimmerman, 2009). Therefore, measuring academic challenges not only aids in understanding students' experiences but also serves as a diagnostic tool for providing timely support to facilitate students' growth and development.

The concept of challenges is essentially subjective. Self-report measures can provide a snapshot of learners' perceptions and experiences. Regardless of potential deviations from behavioral data (Winne & Jamieson-Noel, 2002), individuals' appraisals of challenges and subsequent strategy deployment is based on their self-perceptions. This recognition highlights the significance of incorporating students' subjective viewpoints into research on learning and self-regulation.

Measuring perceived academic challenges can also provide a context for better interpretation of students' learning behaviors and beliefs. For instance, the intention behind a strategy deployment may be better understood by an observer in the context of the type of challenges it was intended to address. Bringing academic challenges into the scene facilitates an understanding of behaviors and beliefs in the context of challenges, offering a more comprehensive view of how students' intention, beliefs, and behaviours interact.

Moreover, measuring academic challenges may empower students to develop metacognitive awareness of their learning struggles. When followed by personalized feedback and resources guiding them how to address these challenges, students can gain insight into a structured understanding of their challenges which can facilitate what they can do to strategically tackle those challenges. In other word, measuring academic challenges can serve as a self-assessment and self-reflective tool (e.g., Panadero et al., 2016, 2017), enabling students to

systematize their recognition of their challenges and devise effective strategies to overcome them.

In summary, measuring academic challenges carries positive implications for both individual learners and institutional diagnostics, while also advancing research. This dual-purpose design, which allows for both research and self-assessment, is predicated on the assumption that the data needed to understand and improve learning is consistent for both researchers and learners (Butler, 2023; Hadwin et al., 2019; Winne & Butler, 1995; Winne, 2017b).

Academic Challenges Scale

The Academic Challenges scale is one of the four measures of the Self-Regulated Learning Profiles and Self-Diagnostic tool (SRL-PSD, Hadwin et al., 2024). The other three measures gauge students' perceptions of the importance of self-regulatory practices (SRL-importance), their confidence in deploying these practices (SRL-self-efficacy), and their self-reported frequency of deploying self-regulatory practices (SRL-Practices). Employing all four measures simultaneously yields a comprehensive and interconnected profile of students' self-regulated learning.

Comprising 31 items distributed across six subscales, the Academic Challenges scale measures the extent to which students have encountered various challenges in their studying over the preceding two weeks.

The Academic Challenges scale underwent iterative development spanning a decade of data collection and pilot testing with post-secondary students. Initially, Hadwin et al. requested students' documentation of challenges encountered during weekly study sessions throughout a semester. These responses were then qualitatively analyzed and structured into a dropdown list.

Additionally, an open text field was provided for contributions of any unlisted challenges. Continuous refinement and expansion of the list occurred until reaching a saturation point, indicated by a negligible rate of uncategorized challenge types reported over two consecutive academic semesters.

Building on this, it was hypothesized that the challenges could be classified into five primary areas of regulation associated with student success: motivation, behavior, cognition, metacognition, and social-emotional challenges. This theoretical model underwent testing via Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) using two samples of university students (Hadwin et al., 2024). As this initial model displayed suboptimal fit, a reconfiguration was undertaken, leading to the emergence of the sixth factor called initiating and sustaining engagement challenges. The current version of the scale comprises challenges related to: (1) motivation, (2) initiating and sustaining engagement, (3) goal and time management (formerly called behavioural), (4) cognitive, (5) metacognitive, and (6) social and emotional challenges. This 6-category model demonstrated adequate psychometric properties. Additionally, the convergent validity of the scale in relation to the other three measures of SRL-PSD, and its predictive validity in predicting end-of-semester GPA, were also established (Hadwin et al., 2024).

In a previous study (Study 2; Rostampour et al., in preparation), a unidimensional graded Item Response Theory (IRT) model was employed to conduct an item-level analysis of the academic challenges scale. This study supported the psychometric adequacy and criterion validity of the overall scale and the subscales as discrete unidimensional constructs.

Additionally, the psychometric adequacy of the composite scores for the overall scale and the

subscales was also examined and supported. However, the multidimensionality of the scale was not investigated in this study.

The previous validation of the academic challenges scale as a multidimensional construct used correlated-factor CFA (Hadwin et al., 2024). The six subscales of academic challenges exhibited moderate to high correlations with one another (Hadwin et al., 2024). Such substantial common variance between the subscales suggest that they may aim to capture distinct dimensions of a broader construct. On the other hand, high correlations might raise concerns regarding discriminant validity, suggesting that the subscales may not be sufficiently distinct. Consequently, within the academic challenges context, such a covariance structure could imply that the subscales collectively represent an overarching construct of general academic challenges.

Considering the multi-faceted and interconnected nature of academic challenges, it is also unreasonable to believe that an indicator of a challenge would exclusively load on a single factor with no contributions received from other factors.

To exemplify, a student's challenge with time management can often be rooted in a combination of cognitive, metacognitive, motivational, behavioral, social, and emotional areas. Cognitively, they may struggle with processing and understanding information efficiently, leading to longer study times and reduced productivity. Metacognitively, difficulties in planning and monitoring their study schedule can result in disorganization and inefficiency in managing time. Motivationally, a lack of intrinsic motivation or low self-efficacy can lead to procrastination and low engagement. Behaviorally, undisciplined study habits and environments can hinder their ability to manage tasks effectively. Socially, the inability to manage social distractions can exacerbate these problems. Emotionally, stress, anxiety, and emotional instability can further disrupt their focus and concentration, making it difficult to maintain a

consistent study routine. These interconnected issues are an example of a complex web of challenges that can impede effective time management. Consequently, it is crucial to enhance the ecological representativeness of the construct by selecting a measurement model that can operationalize this interconnectivity.

The Present Study

The purpose of the current study is to enhance the ecological representativeness of the previous validation of the academic challenges scale as a correlated factor model by addressing and mitigating two of its key potential limitations. The first limitation is the model's discounting of an overarching factor that underlies all responses. The second limitation is the restrictive assumption that each item is influenced exclusively by a single latent factor, suppressing the impact of any other latent factors in the scale on the item.

The first limitation was addressed by modeling an overarching academic challenge factor using second-order and bifactor measurement models. These approaches allow for simultaneous modeling of a set of subordinate factors and one or more superordinate factor(s). The second limitation was relaxed through modeling of cross-loadings by retesting all CFA measurement models using Exploratory Structural Equation Modeling.

In this study, six measurement models shown in Figure 1 were tested and compared to determine the best-fitting model that represents the multidimensionality of academic challenges scale. These models include: Correlated CFA (conducted for revalidation purposes), Second-order CFA, Bifactor CFA, Correlated ESEM, Second-order ESEM, and Bifactor ESEM.

The three models in the first row of Figure 1 represent CFA models and the three models in the second row represent the ESEM versions of those models. In the case of bifactor ESEM

and second-order ESEM measurement models, both of the limitations discussed above are simultaneously addressed.

If a variant of the bifactor model is chosen as the best-fitting model (bifactor CFA or bifactor ESEM), it becomes necessary to investigate the reliability and dimensionality of the construct using a set of ancillary analyses. This is crucial because, in the bifactor approach, general and specific factors compete to explain the variance of items. To further ensure the robustness of the model, beyond indices of psychometric validity, the reliability and internal consistency of both general and specific factors, in relation to each other, will be thoroughly investigated.

The bivariate and multivariate associations of factors across all measurement models will be explored with three criterion variables: GPA expectations, GPA confidence, and end-of-semester GPA. This exploration will be done to understand how different measurement models relate to the criterion variables, providing insight into the predictive validity and practical implications of each model.

Finally, to investigate the stability of the measurement model across different samples, measurement invariance will be tested across three demographic characteristics: year of study (first-year, second-year), sex at birth (male, female), and the two participating cohorts (fall and winter intake). This testing is crucial to determine whether the model behaves consistently and produces comparable results across these diverse groups.

Materials and Methods

Participants and Data Collection Procedure

The data for this study were collected as part of the UniVenture program, a wellness program designed specifically for university students. This program offered specialized workshops that provided students with essential skills, helped them establish long-term goals, and assisted them in using their unique personalities to achieve their goals effectively.

The current study used baseline data collected during the first five weeks of the Fall 2021 and Winter 2022 semesters through the REDCap electronic data capture tools (Harris et al., 2019). The dataset included university students from five Canadian universities across four provinces. GPA data was obtained from the universities' registrar's offices. The sample comprised 2,671 participants, with a distribution of 65% female, 32% male, and 3% non-binary. 56% of participants were first-year students, and 44% were second-year students. The age range of the students was 18 to 25 years, with an average age of 19 and a standard deviation of 1.5.

Data collection during this baseline phase, which formed the basis for this study, included student responses on various topics such as demographics, personality, substance and alcohol use, anxiety, and depression, in addition to academic challenges.

A retrospective power analysis indicated that a model with $\alpha = 0.05$, $df = 269$, and $N = 2671$ would yield acceptable power ($1 - \beta = 0.99$), being satisfied by the current sample size (Jobst et al., 2023) (see supplementary materials for more details).

Measures

The Academic Challenges Scale

The Academic Challenges Scale (supplementary materials, Table 1,) is one of the four measures in the Self-Regulated Learning Profiles and Self-Diagnostic tool (SRL-PSD, Hadwin et al., 2024). The Academic Challenges Scale, consisting of 31 items across six subscales, measures the extent to which students have encountered various challenges in their studying over the preceding two weeks. These factors include: (1) motivation, (2) initiating and sustaining engagement, (3) goal and time management, (4) cognitive, (5) metacognitive, and (6) social and emotional challenges.

The Motivation Challenge factor, encompassing 3 items measures struggles related to self-efficacy, perceived value of work, and persistence, reflecting crucial motivational elements for academic engagement. An exemplar item from this scale is *struggled with believing I can do my work*.

The Initiating and Sustaining Engagement (ISE) scale comprises 4 items that measure the internal motivation to start and persist with tasks, as well as the ability to manage attention. An exemplar item from this scale is *struggled with wanting to do my schoolwork*.

The Goal and Time Management (GTM) scale, defined by 4 items, measures struggles with effective time and task management, organizing time and tasks, keeping commitments or goals, and meeting deadlines. An exemplar item from this scale is *struggled with managing my time, tasks, or goals*.

The Cognitive Challenges Scale (COG), consisting of 6 items, measures struggles with cognitive demands of studying such as processing, retaining, and applying academic content. An exemplar item from this scale is *struggled with remembering things that I learned*.

The Metacognitive Scale (MET) includes 8 items measuring challenges related to planning, monitoring, and adjusting learning approaches. These challenges reflect struggles with higher-order thinking required for task understanding, goal setting, progress monitoring, and strategy application (Winne & Hadwin, 1998). An exemplar item from this scale is *struggled with breaking tasks down*.

The Social and Emotional Scale (SE) comprises 6 items measuring struggles with interpersonal and emotional aspects of the student experience. These challenges include forming connections, maintaining relationships, finding enjoyment in the academic environment, and managing emotional well-being. An exemplar item is *struggled with feeling connected*.

The Likert scale employed for collecting responses was a 5-point scale, encompassing options from strongly disagree (1) to strongly agree (5), with disagree (2), neutral (3), and agree (4) in between. The selection of a two-week timeframe aimed to strike a balance between capturing recent study experiences and ensuring an adequate range of recallable incidents.

Criterion Measures

GPA Expectation. GPA expectation was gauged through a single item that asked students about their anticipated GPA for the current academic semester. This item prompts participants to select the Grade Point Average (GPA) they realistically expect achieving. The response scale ranged from highest to lowest GPA: High GPA, Strong GPA, Satisfactory GPA, Low GPA, and Failing GPA. Each descriptor was accompanied by a site-specific numeric GPA range (e.g., Strong GPA, 2.7 - 3.7).

GPA Self-Efficacy. GPA Self-Efficacy was also assessed by a single item that measured how confident students feel about reaching their expected GPA. Once students have identified

their anticipated GPA, they rated their confidence for achieving their expected GPA with options ranging from not at all confident to extremely confident and a neutral option as the midpoint.

End of Semester GPA. End-of-semester GPAs were collected from the registrar offices of each university. Given the variety of GPA scales used by the institutions, it was necessary to standardize these grades to a common 5-point letter grade scale. This standardization involved converting numerical grades into letter grades according to conversion guidelines provided by each university.

The adoption of a 5-point letter grade scale was driven by its use as the most detailed scale in one of the participating universities. This standardized scale facilitates consistent comparisons across different institutions, where an *A* denotes “Excellent”, a *B* signifies “Very Good”, a *C* is “Satisfactory”, a *D* means “Barely Passed”, and an *E* is marked as “Failed”.

Analytic Strategy

Sequential Strategy for Testing Alternative Measurement Models

In this study, to explore the multidimensional construct of academic challenges, the sequential strategy involving testing alternative first-order CFA, second-order CFA, bifactor-CFA, and first-order ESEM, second-order ESEM, and bifactor-ESEM measurement models was used (see Figure 1; Morin, 2023; Morin et al., 2016; Morin et al., 2020).

Students' responses to the academic challenges scale were first modelled using a unidimensional model, before moving to the six alternative models mentioned. Testing and reporting unidimensional and first-order correlated factors models as initial steps provide useful context to better understand findings from more advanced modeling approaches (Eid, 2020). Criteria to choose and retain a measurement model are described in next sections.

The unidimensional and first-order CFA models aim to reaffirm the psychometric validity of the measure using a larger sample drawn from five Canadian universities. The higher-order and bifactor models, are used to explore the potential existence of an overarching latent construct underpinning academic challenges. As depicted in Figure 1, the measurement models including correlated factor, higher-order and bifactor models will also be implemented within the Exploratory Structural Equation Modeling (ESEM) framework.

In the case of correlated ESEM (model *d*), an item's variance is explained by multiple latent factors simultaneously, without necessarily implying the presence of an unmeasured factor, such as a general factor. The ESEM variants of higher-order and bifactor models are tested to complement the objective of modeling the general factor with incorporating cross-loadings. To elaborate, a bifactor-ESEM model complements the bifactor CFA, by modeling cross-loadings. This model enables modeling a potential general factor, while also relaxing the

strict assumption of local independence by allowing estimation of cross-loadings. That is, the variance of each item is assumed to be decomposable into four sources:

1. The variance attributable to the target specific factor.
2. The variance attributable to the general factor.
3. The variance attributable to non-target factors.
4. The variance unique to the item (residual variance).

The third source of item variance is what a bifactor-ESEM model introduces, beyond what is offered by bifactor-CFA.

Criteria for Evaluating Measurement Models

Selecting an Exploratory Structural Equation Modeling (ESEM) model involves more than just comparing measurement models based on fit indices. Additional criteria must be considered, such as the reduction in correlation relative to the Confirmatory Factor Analysis (CFA) model, the magnitude of cross-loadings, and the primary loadings, all of which shape the interpretation and definition of factors. While the superiority of a first order ESEM is typically compared against a CFA model, the superiority of a bifactor ESEM solution should be evaluated in comparison to both the first order ESEM and bifactor CFA models. These evaluative criteria are summarized in Table 2.

Table 2
Criteria for retaining ESEM Measurement Models

Criteria for retaining a First-Order ESEM Measurement Model

Fit Indices	model fit is improved relative to the CFA model
Factor correlations	factor correlations are reduced in ESEM relative to those of CFA
Cross-loadings	(a) cross-loadings are small to medium (< 0.50) (b) cross-loadings in each factor $<$ primary loadings of items on that factor -In case of minor violation of (a) or (b), a solid theoretic explanation (e.g., unmodelled factor) or methodological (method effects due to similar words or item sequencing) applies.
Primary Factor Loadings	Strong and theoretically meaningful loadings for most of items ($\lambda > .40$ and ideally $\lambda > .50$)

Criteria for retaining a Bifactor ESEM Measurement Model

Theoretical proofs supporting that the constructs follow a global/specific structure	
Fit Indices	model fit is improved relative to the initial CFA and first-order ESEM model
Primary Factor Loadings	the G-factor and S-factors are well-defined by their <i>a priori</i> factor loadings ($\lambda > .30$, ideally $\lambda > .50$ on at least one of the G- or S-factors) ^{1,2}
cross-loadings	Cross-loadings are reduced relative to the first order ESEM model or at least remain non-problematic.
Measures of Reliability	the reliability estimates (i.e., ω) of the ESEM should be acceptable. See the reliability section.

Note. Reference: Morin et al. (2020); van Zyl & ten Klooster, (2021)

¹ In bifactor models, the reliable item variance is partitioned between the G-factor and S-factors, resulting in factor loadings that tend to be smaller than those observed in first-order and higher-order models.

² It should also be noted that while support for a bifactor solution requires that at least some S-factors retain specificity, bifactor solutions are known to be robust even when S-factors collapse. This phenomenon indicates that the items associated with these subscales primarily serve to define the G-factor and retain limited specificity once this G-factor has been accounted for.

Criteria for Evaluating the Fit Indices

The criteria for evaluating fit indices in this study are provided in Table 3. Since the response scale used for the academic challenges scale is a 5-point Likert scale, a Weighted Least Squares Mean and Variance adjusted estimator (WLSMV) is employed (Li, 2016). In addition to

using WLSMV as the main estimator, all measurement models were also tested using Maximum Likelihood Estimation with Robust Standard Errors (MLR). This estimator was used to compute information criteria such as AIC, BIC and aBIC and to compare indices with those of WLSMV.

In evaluating the misfit of the models, greater emphasis was placed on SRMR over RMSEA, as literature indicates that SRMR generally outperforms RMSEA when modeling categorical data (Shi et al., 2020). This is because RMSEA penalizes model complexity; therefore, RMSEA values obtained from models estimated by WLSMV should be evaluated with greater leniency compared to those derived from the ML estimator (van Zyl & ten Klooster, 2021).

Table 3
Evaluation Criteria for Interpreting Fit Indices

Index	Evaluation Criteria
Chi-Square (χ^2)	- Lowest comparative χ^2 value - Non-Significant χ^2 value ($p > 0.01$) ¹
comparative fit index (CFI)	CFI $\geq .95$, excellent fit CFI $\geq .90$, acceptable fit
Tucker-Lewis Index (TLI)	TLI $\geq .95$, excellent fit TLI $\geq .90$, acceptable fit
Root Mean Square of Error Approximation (RMSEA)	$RMSEA \leq .05$ $RMSEA \text{ 90\% upper CI } \leq 0.08$ $RMSEA \leq .08$ } excellent fit acceptable fit
Standardized Root Mean Square Residual (SRMR)	SRMR $\leq .05$ excellent fit SRMR $\leq .08$ acceptable fit
Akaike Information Criterion (AIC) Bayesian Information Criterion (BIC) Sample-Size Adjusted BIC (aBIC)	Lowest comparative information Criterion value

Note. ¹The Chi-Square (χ^2) test is highly sensitive to sample size, with its significance often influenced by the number of observations (N). This sensitivity can lead to misleading significance levels, especially in large samples, where even minor deviations from expected distributions might appear statistically significant.

References: (Byrne, 2013; Hu & Bentler, 1999; Marsh et al., 2005; Marsh et al., 2014; West et al., 2023).

Evaluating the Criterion Validity

Criterion validity is crucial as it involves judging how well one variable or set of variables is associated with an outcome, serving as a practical indicator of the usability of a measure. Testing for criterion validity is essential because it ensures that the instrument is associated with practical outcomes, thus confirming its effectiveness in real-world settings.

In this study, three criterion variables were used: (a) students' expected GPA (EGPA), (b) their self-efficacy in achieving that GPA (GPA_SE), and (c) their actual registrar end-of-semester GPA (GPA).

To examine the association between academic challenges and criterion variables, the ESEM-within-CFA approach was employed. This approach transforms an ESEM model into a CFA structure while maintaining the same chi-square, degrees of freedom, fit statistics, and parameter estimates so that it can be used in complex structural path analyses (c.f., Morin et al., 2013). In this method, non-standardized loadings from the ESEM model are used as starting values for the item loadings in the ESEM within CFA model. Factor variances are constrained to one, and one item per latent factor is fixed to its original ESEM item loading (Morin & Asparouhov, 2018; Morin et al., 2016). Consequently, the ESEM-within-CFA framework allows researchers to effectively use ESEM models for complex research questions.

In this study, the association between academic challenges and criterion variables was examined across all measurement models using two methods: (a) bivariate associations between factors and criterion variables, and (b) multivariate associations where all factors of a measurement model concurrently predicted a criterion variable.

The rationale for examining associations across all measurement models is to explore how different modeling approaches impact the relationships between the factors and criterion

variables. For example, the study examines how including cross-loadings in ESEM, overarching factors in second-order and bifactor models, and combinations of these approaches in bifactor ESEM affected the relationships between the factors and criterion variables. This comprehensive analysis was conducted because all measurement models achieved at least an acceptable level of fit.

The bivariate/multivariate approach was employed to discern the suppressive and additive effects of factors when associated with the criterion variables. As discussed earlier, the orthogonality of specific factors and the general factor in bifactor models allows for the proper disaggregation of these components, enabling their simultaneous use in a structural model. This characteristic allows for examination of the effects of general and specific factors above and beyond each other in predicting the criterion outcomes. However, as previously explained, the concurrent modeling of the second-order factor and their subordinate factors within a second-order factor model is not feasible due to the overlapping nature of these components.

Because students rated their GPA self-efficacy with respect to their expected GPA (EGPA), EGPA was included as a covariate in the model assessing the association between academic challenges and GPA self-efficacy. This means that the relationship between GPA self-efficacy and academic challenges was examined controlling for variations in students' GPA expectations.

Reliability of the Bifactor Measurement Model

It is important to note that incorporating a general factor in bifactor models or cross-loadings in ESEM models does not necessarily enhance the instrument's reliability. Despite potential improvements in model fit, researchers should meticulously examine the reliability parameters of these models (Reise et al., 2013; Rodriguez et al., 2016). Evidence suggests that

traditional goodness-of-fit indices may statistically favor bifactor models (Bonifay et al., 2017, Gignac, 2008, 2016; Morgan, et al., 2015). In this context, it is crucial to use a set of auxiliary measures for a more comprehensive evaluation of the quality of a bifactor model. These measures facilitate the examination of the dimensionality and reliability of measurement models and include omega reliability coefficients (ω , ω_S , ω_H , ω_{HS}), factor determinacy (FD), construct replicability (H), explained common variance (ECV_{GS} , ECV_{SG} , ECV_{SS} , I_{ECV}), the percentage of uncontaminated correlations (PUC), and the average relative parameter bias ($ARPB$).

In a bifactor model, item variance and the covariance between items can be attributed to multiple common sources, such as the general factor (g), specific factors (s), and uniqueness (random error and systematic bias). These auxiliary indices enable the partitioning of variance among these sources, enhancing our understanding of the scale's multidimensionality (Simms, et al., 2008). By examining these indices, researchers can address questions regarding the nature of multidimensionality of the measure, the reliability of the general and specific factors, (Chen et al., 2006; Raykov & Pohl, 2013; Raykov & Marcoulides, 2018), and the adequacy of total and subscales scores for the measure. These indices assess various aspects of psychometric characteristics of the measure. Indices such as ω are used for evaluating the quality of unit-weighted total and subscale score composites. Indices like FD (factor determinacy) are used for assessing the quality of factor score estimates that can be used as proxies of the true factors. Additionally, indices such as H (construct replicability) are used to evaluate the quality of measurement models within structural equation modeling.

It is essential to recognize that splitting the reliable item variance into the G-factor and S-factors in bifactor models makes both factor loadings and reliability estimates in bifactor models

to be lower than those seen in higher-order models. Therefore, more lenient criteria have been recommended for evaluating the composite reliability of S-factors in bifactor models, particularly because S-factors often involve fewer items and lower loadings. As an example, Morin et al. (2020) suggests using a threshold closer to $\geq .50$, rather than the conventional $\geq .70$ to assess the reliability of the specific factors.

The auxiliary indices used in this study to evaluate the reliability of the bifactor models are briefly defined in Table 5. The criteria used to evaluate these indices appear in the third column. When multiple criteria are mentioned for an index, the most conservative one was selected as the criterion for evaluating the dimensionality and reliability of the academic challenges scale.

Table 5
Ancillary Measures of Reliability for the Bifactor Model

Index	Definition	Cut-off	References
Total Omega (ω)	the proportion of variance in the total observed score that can be attributed to all modelled sources of common variance (the general factor and all specific factors)	The heuristic rule of thumb: >.7 acceptable >.8 good Reliability >.9 excellent	(Flora 2020; McDonald, 2013; Reise, et al, 2013; Zinbarg et al., 2005)
Subscale Omega (ω S)	the proportion of variance in scores on a specific subscale that can be attributed solely to that specific factor	The same cut-of as the above. However, ω S is sensitive to the number of items with lower items resulting in smaller internal consistency values.	(c.f. Dunn, et al., 2014; Lance, et al., 2006), for a contrasting viewpoint about the cut-off values of measures of internal consistency
Hierarchical Omega (ω H)	the proportion of variance of the total scores that can uniquely be explained by the general factor alone. (Treating variability in scores due to specific factors as measurement error)	ω H > .80 would indicate that the general factor is the main source of systematic variance compared to the influence of the specific factors. In this case, the scores can be considered essentially unidimensional.	(Rodriguez, et al., 2016; Flora, 2020)
Hierarchical Omega for Subscale (ω HS)	the proportion of variance of a subscale score after separating the variability attributed to the general factor	ω HS $\geq$.30 substantial .20 $\leq$ ω HS <.30 moderate ω HS < .20 low	(Reise et al., 2010; Reise, et al., 2013a; Reise, et al., 2013b; Smits et al., 2014)
Percentage of Reliable Variance (PRV)	This index is the ratio of ω H to ω ; and therefore, it can be conceptually understood as the percentage of the total reliability that can be attributed to the reliability of the general factor	PVR > 50 indicate that half of the reliable variation in test score is produced by the general factor (or the specific one, if the index is calculated as the ratio of ω HS to ω)	(Bonifay, et al., 2017; Reise, et al., 2013a; Rodriguez, et al., 2016)
Explained Common Variance (ECV)	the proportion of the common variance that can be attributed to the general factor	provisional cut-off points: ECV> .70 or ECV>.8 suggests that the common variance among the specific factors is small compared to the general factor; and therefore, that the data would fit an essentially unidimensional model	(Reise, 2013b; Rodriguez et al., 2016; Sijtsma, 2009; Stucky, & Edelen, 2014; Ten Berge & Sočan, 2004)
ECV_GS	the proportion of common variance of the items in each specific factor which is due to the general factor		(Dueber, 2017; Stucky, & Edelen, 2014)
ECV_SG	the proportion of common variance of all items which is due to the specific factor		(Dueber, 2017; Stucky, & Edelen, 2014)
ECV_SS	the proportion of common variance of the items in each factor which is due to that factor		(Dueber, 2017; Stucky, & Edelen, 2014)
I-ECV	the proportion of true variance of each item that is explained by the general factor.	I-ECV> .85 suggests a strong influence of the general factor on the variance of the item	(Dueber, 2017; Stucky, & Edelen, 2014)
Percentage of Uncontaminated Correlations (PUC)	the percentage of the variance in pairwise item correlations that is attributable solely to the general factor, excluding the effects of specific factors and error variances on those correlations (the percentage of the total correlations unaffected by multidimensionality)	the PUC together with the ECV provide information about the bias towards forcing multidimensional data into unidimensional models. the scale can be treated as if it were unidimensional under one of these conditions: PUC> .80, the ECV value is less important. PUC< .80, ECV>.60 ECV>.70, PUC>.70	(Bonifay et al., 2017; Reise, et al., 2013b; Rodriguez et al., 2016)
Factor Determinacy (FD)	quantifies the viability of using factor scores as proxies for individual differences on a latent variable. Can be interpreted as the correlation between factor scores and factors	Possible values range from 0 to 1, with values closer to 1 indicating better determinacy. Factors score estimates only should be considered in research if FD> .80 or FD>.90.	(Beauducel, 2011; Grice, 2001; Gorsuch, 2014)
Construct Replicability (H)	provides an assessment of the robustness or replicability of the construct in SEM measurement models across different samples. quantifies how much variance in the total scores is explained by the latent factor compared to the total variance which is unexplained by the factor.	For the general factor: H>.70 suggest replicability of the general factor Sensitive to the number of items and the magnitude of factor loadings	(Hancock & Mueller, 2001; Rodriguez et al., 2016)
Average Relative Parameter Bias (ARPB)	is a measure for examining the difference between the factor loading of a unidimensional model and the general factor loading of the bifactor model	Maximum of ARPB10-15% may be acceptable and poses no serious concern	(Muthén, et al., 1987; Rodriguez, et al., 2016)

Note.

Evaluating Measurement Invariance

Testing measurement invariance as part of construct validation is crucial because it ensures that a measurement instrument operates equivalently across different groups or conditions. By confirming that the same items hold the same psychometric properties for all specified groups, measurement invariance enhances confidence that comparisons of the construct across these groups are valid and meaningful. In this study, measurement invariance was assessed across sex at birth (male $N=969$, female $N=1698$), participating cohorts (Fall intake $N=1246$, Winter intake $N=1427$), and year of study (first year $N=1536$, second year $N=1137$).

To ensure measurement invariance across groups, multigroup confirmatory factor analysis (MGCFA) was employed. Following the guidelines suggested by Brown (2015), the seven steps summarized in Table 4 were used to ensure measurement invariance.

Table 4
Steps for Testing Measurement Invariance

Step	Description
0. Whole Sample Model	The baseline model was tested on the entire sample to ensure a good overall fit before proceeding with multigroup analyses.
1. Separate Model Testing	The model was tested separately for each group to establish a baseline comparison.
2. Configural Invariance	A configural invariance step was implemented to ensure that the same pattern of fixed and free factor loadings was specified for each group.
3. Metric Invariance (Weak)	A metric invariance model was estimated, constraining the factor loadings (λ) to be equal across groups to test if the construct is measured in the same way for both groups.
4. Scalar Invariance (Strong)	A scalar invariance model was estimated, where both item thresholds (ν) and factor loadings (λ) were constrained to be invariant across groups, testing if the item intercepts are equivalent.
5. Strict Invariance	A strict invariance model was estimated, in which the factor loadings (λ), item thresholds (ν), and item uniquenesses (δ) were constrained to be invariant across groups, ensuring that both the measures and the measurement error are equivalent.
6. Latent Variance-Covariance Invariance	The model was further constrained so that the variance-covariance parameters (δ , ξ/ϕ) of the latent variables are equal across groups, to test if the constructs have the same variance and inter-correlations.
7. Latent Mean Invariance	In addition to the constraints from previous steps, the latent means (η) are set to be equal across groups, assessing whether the average scores on the latent variables are the same for each group.

The initial six steps (0-5) in testing measurement invariance aim to identify potential measurement biases concerning the nature, structure, and reliability of the construct. The last two steps, which focus on variance, covariances, and means, expand the analysis to probe for group differences in these areas. Establishing measurement invariance at the levels of latent variance-covariance (Step 5) and latent means (Step 6) is a prerequisite for making valid comparisons of systematic differences between groups (Marsh et al., 2013).

In assessing model fit for the measurement invariance tests, the DIFFTEST function in Mplus was initially used to evaluate fit improvement through the $MD\Delta\chi^2$ index (Asparouhov et al., 2006). However, due to the oversensitivity of $MD\Delta\chi^2$ to sample size, additional indices were employed. Specifically, the criteria for changes in the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA) proposed by Chen (2007) were used. Changes in CFI (ΔCFI) of less than or equal to 0.010 and changes in RMSEA ($\Delta RMSEA$) of less than or equal to 0.015 were considered indicative of invariance across groups. This approach recognizes the need to complement chi-square difference tests with other fit indices when comparing nested models (Chen, 2007; Cheung & Rensvold, 2002). The overall fit of the constrained models was assessed using the same criteria applied to the unconstrained models, as detailed earlier in this section. While earlier literature indicated that conventional guidelines for model fit could be applied to WLSMV estimation (Yu, 2002), recent findings advocate for less stringent criteria when using this estimator compared to variant of the ML estimator (Shi et al., 2018; Xia & Yang, 2019).

Results

Descriptive statistics are provided in the supplementary materials (Table B). Considering that, on average, less than 3 percent of the data were missing and the estimator (i.e., WLSMV) demanded prior handling of missing data, pairwise deletion was employed for the analyses.

The data analysis was conducted using Mplus version 8.3. (Muthén & Muthén, 1998-2021). For generating codes for the ESEM within CFA models, the ESEM code generator for Mplus (De Beer & Van Zyl, 2019) was used. The ESEM invariance syntax generator for Mplus (De Beer & Morin, 2022) was used to produce codes for the measurement invariance testing. The ancillary reliability analysis was computed using the R package BifactorIndicesCalculator (v. 0.2.2; Dueber, 2022), and the power analysis was conducted using the R package semPower (v. 2.1.0; Moshagen & Bader, 2023).

Conceptualization

As discussed in the analytic strategy section, to explore the factorial properties of the academic challenges scale, a competing measurement modeling strategy was employed comparing variations of CFA with ESEM models. The measurement models shown in Table 6 were estimated and compared.

Table 6
The Sequential Analytic Strategy Involving Seven Measurement Models

Model 0	Unidimensional CFA Model of Overall Academic Challenge
Model 1	Correlated First-Order CFA Model comprised of Six factors (Figure 1a)
Model 2	Hierarchical CFA Model compromising a second-order factor of academic challenges as well as six first-order factors (Figure 1b)
Model 3	Bifactor CFA Model compromising a general factor as well as six specific factors of academic challenges (Figure 1c)
Model 4	Correlated First-Order ESEM Model comprised of Six factors (Figure 1d)
Model 5	Hierarchical ESEM Model compromising a second-order factor of academic challenges, made up of six first-order factors (Figure 1e)
Model 6	Bifactor ESEM Model compromising a general factor as well as six specific factors of academic challenges (Figure 1f).

Fit Indicators

The measures of fit for the measurement models are reported in Table 7. While the unidimensional model demonstrated poor fit, the other CFA measurement models (Models 1 to 3) achieved marginally acceptable fits, with the 6-factor model (Model 2) exhibiting a slightly better alignment with the data. Examining the fit indices for the ESEM models (Models 4 to 6) revealed that these models all surpassed conventional cut-off values of acceptable fit. Notably, the Bi-factor ESEM (Model 6) showed superior fit and lower misfit according to the indices. It is important to note that, except for the RMSEA values, the other fit indices for all the ESEM models were very strong. As discussed in the analytic criteria section, when dealing with categorical data, greater emphasis should be placed on SRMR, because RMSEA penalizes model complexity, and RMSEA values for models estimated by the WLSMV should be evaluated with greater leniency compared to those derived from the ML estimator (Shi et al., 2018; Xia & Yang, 2019).

Table 7
Measures of Fit

Model	Type	χ^2	df	χ^2/df	RMSEA	90% lower CI	90% Upper CI	CFI	TLI	SRMR	Meet Criteria
<i>Confirmatory Factor Analytical Models</i>											
Model0	Unidimensional	25168.845*	434	57.99	0.146	0.144	0.148	0.795	0.781	0.075	NO
Model1	6-factor	9213.445*	419	21.99	0.089	0.087	0.09	0.927	0.919	0.044	Marginal
Model2	Second-order	10457.790*	428	24.43	0.094	0.092	0.095	0.917	0.91	0.05	Marginal
Model3	Bifactor	10203.526*	403	25.32	0.095	0.094	0.097	0.919	0.906	0.047	Marginal
<i>Exploratory SEM Analytical Models</i>											
Model4	6-factor	4329.487*	294	14.73	0.072	0.07	0.074	0.967	0.947	0.02	YES
Model5	Second-order	4121.686*	303	13.60	0.069	0.067	0.071	0.968	0.952	0.021	YES
Model6	Bifactor	3386.603*	269	12.59	0.066	0.064	0.068	0.974	0.955	0.017	YES

To compare AIC, BIC, and adjusted BIC estimates, the same measurement models were estimated using the MLR estimator (refer to the supplementary materials). The upper 90% CI for RMSEA values were below .05 for all the ESEM models estimated by MLR, underscoring their

strong fit to the data. The superiority of Model 6 (i.e., the bifactor ESEM) was further indicated by its lower AIC, BIC, and aBIC values.

Comparison of the first-order correlated CFA with the first-order correlated ESEM ($\Delta\text{CFI}=0.04$, $\Delta\text{RMSEA}=-0.017$), second-order CFA with second-order ESEM ($\Delta\text{CFI}=0.051$, $\Delta\text{RMSEA}=-0.025$), and bifactor CFA with bifactor ESEM ($\Delta\text{CFI}=0.048$, $\Delta\text{RMSEA}=-0.029$) indicated that overall, ESEM models showed noticeable greater fit than their ICM-CFA equivalent models. Since the CFA models are nested within their ESEM equivalents, a mean and variance adjusted Chi-square test (using Mplus DIFFTEST function) was employed to compare these models. The *p*-values from the three separate tests—to compare the first-order correlated CFA with the first-order correlated ESEM, second-order CFA with second-order ESEM, and bifactor CFA with bifactor ESEM—were significant, highlighting the superiority of the ESEM models over the ICM-CFA models.

The Bifactor ESEM model (model6) fit the data slightly better than the first-order ESEM model, ($\Delta\text{CFI}=0.007$, $\Delta\text{RMSEA}=-0.006$).

In summary, across all models tested, the bifactor ESEM (Model 6), accounting for a global measure of academic challenges as well as cross-loadings, was selected as the best fitting model. The following sections compares the competing models in terms of the other criteria discussed in the analytical strategy section.

Factor Correlations

A comparison of the bivariate correlations between the six factors shown in Table 8 showed that the correlations between factors were significantly reduced in the first order ESEM model compared to the first-order CFA model. Table 8a and 8b demonstrate the correlation matrix between the factors for the CFA and ESEM, respectively while the 8c shows the reduction of correlations (correlations in model 1 MINUS correlations in model 4). Such a reduction of correlation supported that the ESEM model corrected for the positive bias in factor correlation in the CFA model and should be retained.

Table 8*Inter-Factor Correlation for the CFA and ESEM models*

<i>a) Correlated-Factor CFA (Model1)</i>						
	MOT	ISE	GTM	COG	SE	MET
MOT	1					
ISE	0.683	1				
GTM	0.815	0.849	1			
COG	0.756	0.536	0.681	1		
SE	0.669	0.662	0.649	0.607	1	
MET	0.783	0.663	0.835	0.794	0.681	1
<i>b) Correlated-Factor ESEM (Model4)</i>						
MOT	1					
ISE	0.167	1				
GTM	0.242	0.254	1			
COG	0.462	0.202	0.424	1		
SE	0.412	0.414	0.369	0.558	1	
MET	0.427	0.348	0.493	0.671	0.625	1
<i>c) Differences in Correlations Between CFA and ESEM Models Using Fisher's r-to-z Transformation)</i>						
MOT	1					
ISE	0.582	1				
GTM	0.714	0.759	1			
COG	0.452	0.375	0.361	1		
SE	0.355	0.342	0.368	0.074	1	
MET	0.535	0.409	0.581	0.263	0.097	1

Item Uniqueness

The variance estimates were all positive and no anomalies were observed. The R^2 and residual variances for the items across the measurement models are reported in the supplementary materials. On average, the latent factors explained 63, 66, 62, and 67 percent of variability in the items for the ICM-CFA, ESEM, Bifactor CFA, and Bifactor ESEM, respectively. As CFA models are nested in ESEM models, one should not expect to have higher explained variance in a CFA model compared to its ESEM version. The R^2 estimates indicate that failing to model the cross loadings leads to about 5% (67% vs. 62%) less variability in item

variances accounted by the bifactor model. The standardized residual variance for the items ranged from .25 to .43 in the bifactor ESEM model.

Primary Factor Loadings and Cross Loadings

Item-factor loadings across first-order correlated models and bifactor models are reported in the supplementary materials. The factor loadings for the bifactor ESEM model are shown in Table 9.

Table 9
Item-Factor Loadings for the Bifactor ESEM Model

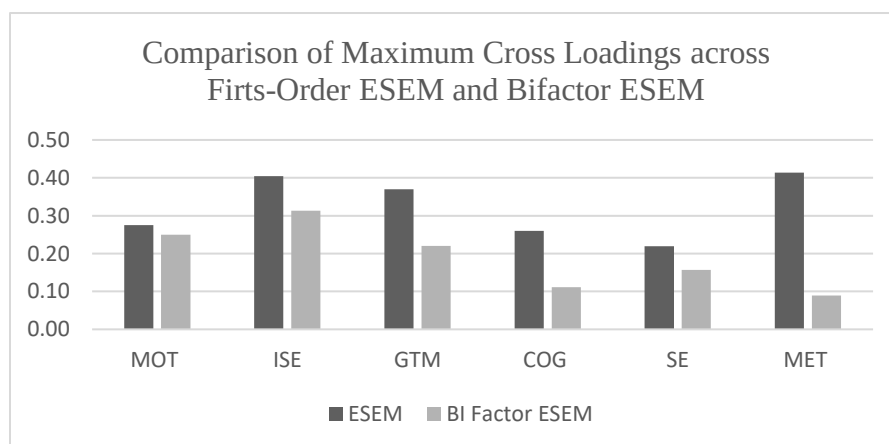
	G	MOT	ISE	GTM	COG	SE	MET
MOT1	0.69	0.36	0.02	0.02	0.03	0.04	-0.08
MOT2	0.60	0.44	0.01	0.08	0.11	0.07	0.02
MOT3	0.69	0.37	0.04	0.13	0.10	0.04	0.01
ISE1	0.56	0.25	0.43	0.16	-0.01	0.05	0.02
ISE2	0.40	0.05	0.30	0.22	-0.04	-0.01	0.06
ISE3	0.53	-0.08	0.55	0.04	-0.03	0.11	-0.01
ISE4	0.64	-0.09	0.56	0.01	-0.11	0.16	0.03
GTM1	0.73	0.19	0.08	0.36	0.06	0.06	0.03
GTM2	0.70	0.00	0.31	0.41	-0.09	-0.01	-0.01
GTM3	0.64	0.14	-0.19	0.38	0.06	-0.11	0.03
GTM4	0.74	-0.14	0.31	0.24	-0.15	-0.01	0.09
COG1	0.67	0.02	0.04	0.04	0.59	0.02	0.04
COG2	0.65	-0.03	0.04	0.01	0.65	0.00	0.03
COG3	0.60	0.05	-0.32	0.13	0.40	0.05	0.01
COG4	0.70	0.05	-0.25	0.06	0.41	0.00	0.00
COG5	0.65	0.05	0.24	-0.21	0.34	0.04	-0.10
COG6	0.68	0.12	0.01	-0.17	0.38	-0.04	-0.10
SE1	0.59	0.11	0.11	-0.11	0.09	0.43	0.07
SE2	0.37	0.11	0.09	-0.06	0.02	0.43	0.08
SE3	0.51	0.15	0.07	-0.06	0.09	0.43	0.06
SE4	0.59	-0.06	0.06	0.05	-0.03	0.65	-0.08
SE5	0.56	-0.05	-0.11	0.08	0.04	0.50	-0.05
SE6	0.61	-0.07	0.13	0.01	-0.10	0.56	-0.10
MET1	0.83	-0.10	-0.07	0.03	-0.05	0.02	-0.16
MET2	0.84	-0.07	-0.18	-0.05	0.07	-0.03	-0.26
MET3	0.86	-0.04	-0.10	-0.10	0.00	-0.05	-0.23
MET4	0.84	-0.07	-0.01	-0.10	-0.02	-0.03	-0.05
MET5	0.77	-0.08	-0.15	-0.03	-0.02	-0.08	0.21
MET6	0.86	-0.07	-0.08	-0.03	-0.01	-0.07	0.24
MET7	0.77	-0.06	0.01	0.07	-0.06	-0.09	0.32
MET8	0.83	-0.05	0.03	-0.04	0.07	-0.01	0.17

Note. The highlighted loadings indicate the primary loadings of items in their target factor. The bold estimates were statistically significant.

Cross Loadings

Maximum of cross loadings within each factor ranged from .22 to .41 (avg=.32) in the correlated ESEM model and from .09 to .31 (avg=.19) in the bifactor ESEM model, indicating that the average cross loadings were lower in the bifactor ESEM model compared to the first-order ESEM model (Figure 2). With the exception of GTM4, cross loadings within factors were weaker than the primary loadings in both the first-order and bifactor ESEM models. This observation indicated that the estimation of cross-loadings did not interrupt the interpretation of factors.

Figure 2



Primary Loadings on Factors

The comparison of average primary factor loadings across measurement models are represented in Figure 3. Such an examination revealed several key points summarized in Table 10.

Figure 3

Comparison of Average Standardized Primary Factor Loadings

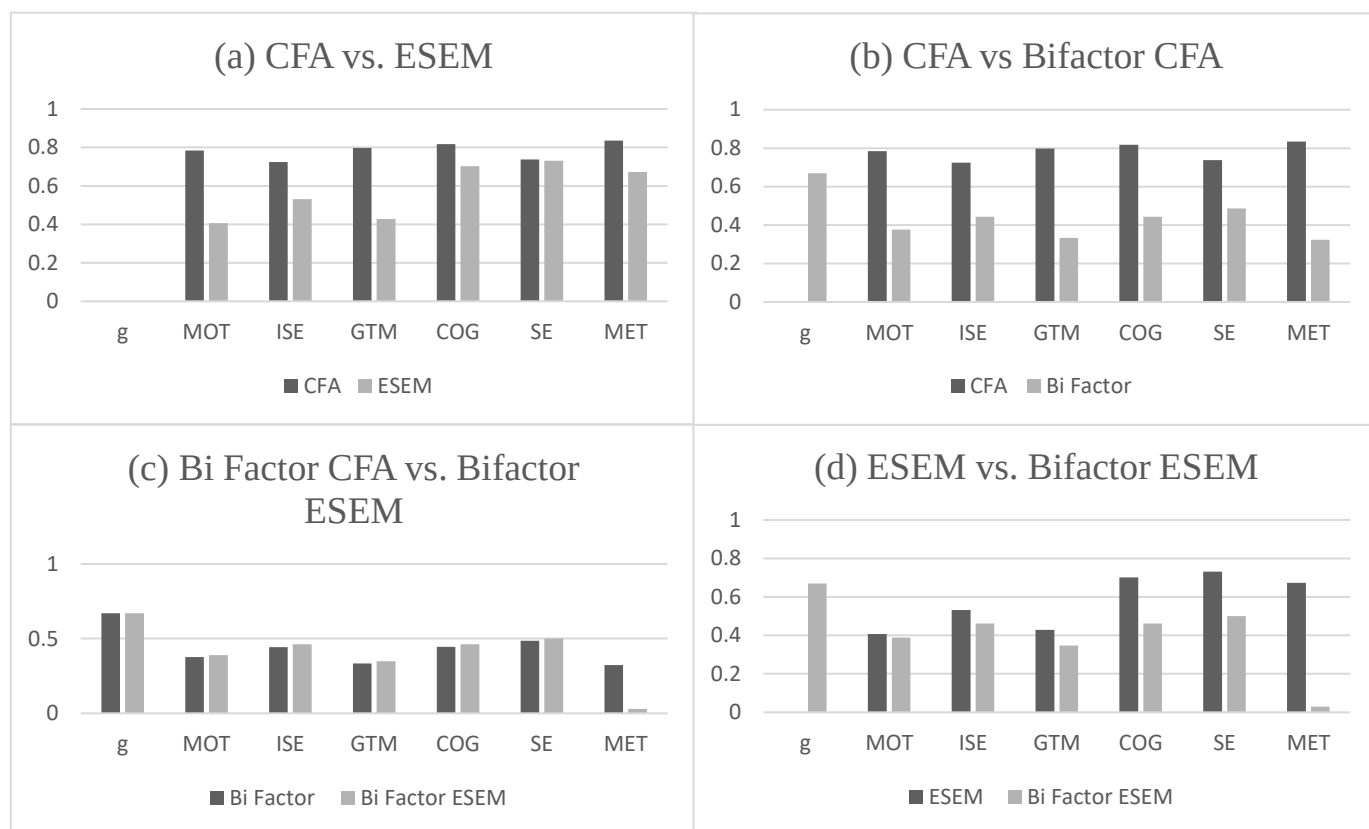


Table 10*Findings from Examining Primary Factor Loadings*

Figure 3a. Comparing average primary factor loadings between CFA and ESEM models	On average, the loadings in the ESEM model are lower than those in the CFA model, particularly for the MOT, GTM, and ISE scales. The higher loadings in the CFA model were interpreted as positive biases, resulting from the forced exclusion of non-target factors' contributions to item variances, which was addressed in the ESEM model.
Figure 3b. Comparing average primary factor loadings between CFA and Bi Factor CFA models	Once the g factor is added to the CFA model, there is a significant reduction in the contribution of the specific factors to item variances. Item loadings on the g factor exceeds their loadings on the specific factors.
Figure 3c. Comparing average primary factor loadings between Bi Factor CFA and Bifactor ESEM models	A comparison between these models showed that the factor loadings were similar, except for the metacognitive scale, which has minimal primary factor loadings in the bifactor ESEM model. (To address the convergence issue caused by linear dependency between items GTM2 and GTM4, their residual variances were allowed to covary in the Bifactor CFA.)
Figure 3d. Comparing average primary factor loadings between ESEM and Bifactor ESEM models	Comparison revealed that the addition of the g factor to the ESEM model significantly reduced the contribution of specific factors to the items' variances, especially in the COG, SE, and MET scales. The average loadings on the metacognitive challenge scale were very low that this subscale was practically vanished in the bifactor ESEM model.

To sum up, in the bifactor ESEM model, all items statistically significantly loaded on the g factor (ranging from .37 to .86, $avg = .67$, $sd = .13$, $p < .001$). Except for item MET4, all the items significantly loaded on their target factor (.30 to .65, $p < .001$), while their magnitude of their loading on their target factors were lower compared to the models that did not include g. Almost

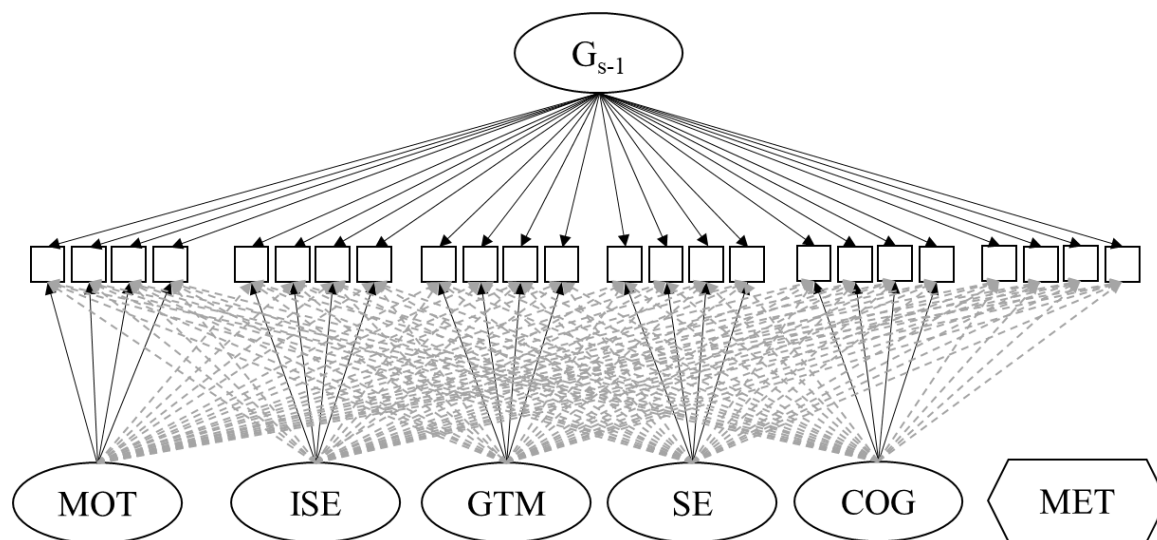
all items loaded more strongly on the g factor than on their specific factors. The only exception was item GTM4 which loaded strongly on the g factor (.74) but its loading on ISE was slightly higher than its loading on its target factor GTM (.30 vs. .24). Such a relative ratio of general to specific loadings across items was considered as the evidence of the necessity of modeling the g factor.

More importantly, the first 4 items of the metacognitive scales were inversely loaded on this scale (Table 9). Looking at their factor loadings on the g factor showed that they had the greatest magnitude of association with the g factor (.83 to .86). This anomaly was resolved through modeling a S-1 bifactor ESEM model which will be described in the next section.

Bifactor S-1 Measurement Model

When factors are expected to contribute asymmetrically to the general factor, meaning that specific factors vary in their contributions to the G-factor, the bifactor S-1 model (see Fig. 4) provides an alternative to the more common symmetrical bifactor model. In a symmetrical bifactor model (e.g., Figure 1b), one S-factor per subscale is estimated, whereas in a bifactor S-1 model, one S-factor is chosen as a reference factor to define the G-factor directly, and S-1 specific factors are estimated. Aside from omitting one S-factor, other specifications of a regular ESEM model are applied (Eid et al., 2017).

For instance, if the metacognitive challenges scale is selected as the reference domain, the G-factor is primarily defined by items from this scale. In other words, the general factor is anchored in the metacognitive challenges. Conceptually, the G-factor predominantly derives its meaning from metacognitive items, with decreasing influence from other items as indicated by their relative factor loadings on the G-factor.

Figure 4*Bifactor S-1 ESEM Mode*

Note. In a bifactor S-1 model, a subset of items (e.g., Metacognitive scale as a reference domain) load exclusively on the general factor. The remaining items (nonreference domains) load on the general factor and their target specific factor.

As items in metacognitive scale loaded strongly on the g factor and minimally on their target factor, the specific factor for the metacognitive challenges items was removed, resulting in a measurement model with one g factor and five specific factors (Figure 4). This means that the g factor became primarily anchored in metacognitive challenge scale. Conceptually, the g factor primarily derived its meaning from the metacognitive challenge items, with decreasing contributions from other items based on their factor loadings on the g factor.

Because the bifactor S-1 model is nested within the symmetrical bifactor-ESEM model (Hand & Lonigan, 2021), one should not expect the former model to exhibit higher fit indices than the latter. In fact, the S-1 model exhibits the same level of global fit as the first order ESEM. The changes in the fit indices were indicative that the bifactor S-1 should be retained

($\Delta\text{CFI}=0.007$, $\Delta\text{RMSEA}=-0.006$). For clarity and to distinguish it from the g in the symmetrical bifactor model, the general factor in the S-1 model is referred to as g_{s-1} .

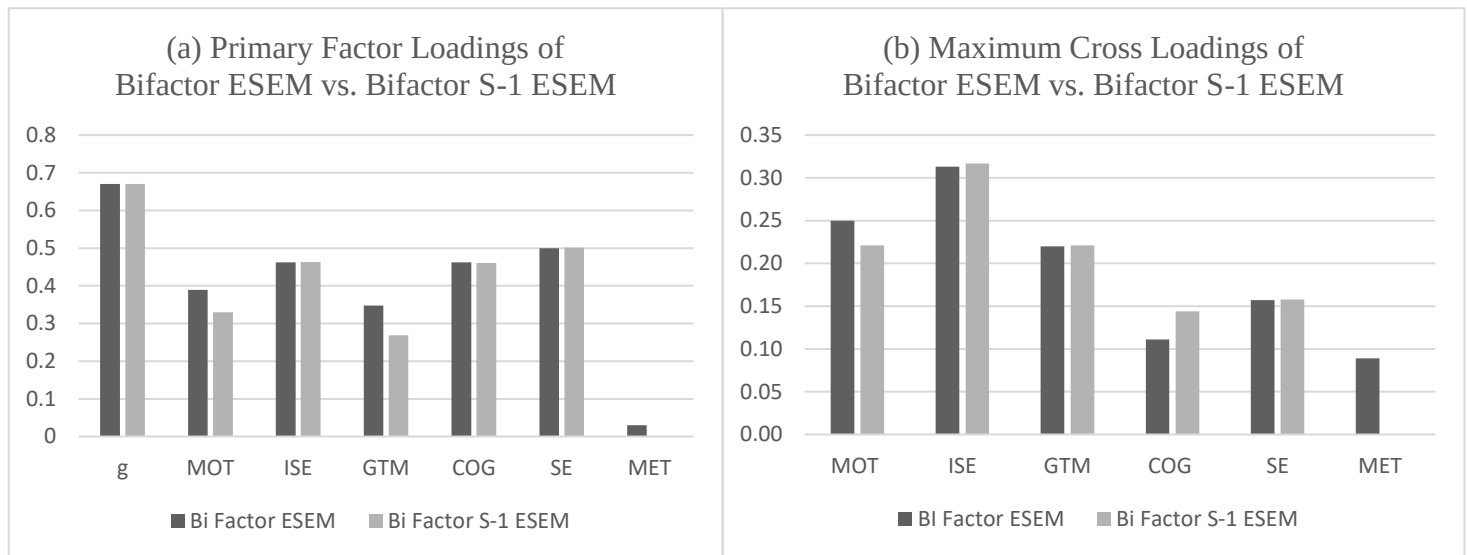
Looking at the factor loadings of the S-1 model (Table 11) and comparing to those of the symmetrical bifactor showed relatively consistent loadings for items both in the general and the specific factors (Figure 5a). All items loaded significantly on their target factors with the anomalies of factor loadings on the metacognitive model solved, as this factor was exclusively captured by the g_{s-1} factor (Table 11). The maximum cross loadings of items within each scale (Figure 5b) ranged from .14 to .32 (avg=.21) and the primary loadings on the factors exceeded all cross loadings. The only two exceptions were the items GTM2 and GTM4. While they were strongly associated with the g_{s-1} factor (.72, .75) their factor loadings on the initiating and sustaining engagement scale were slightly higher than their loadings on the goal and time management factor to which they belonged (.32 vs .27 for GTM2 and .31 vs .18 for GTM4). These items measure the struggles with measuring time and task and when the influence of g_{s-1} on them is controlled, their revisualized variance tended to load more strongly on the ISE rather than on their original factor¹.

¹ These items feature similar wording, which indicates a method effect, resulting in significant residual covariance beyond the influence of the GTM factor. Allowing the residual variance to covary across the items makes the loadings of the GTM2 slightly higher on its target factor than on ISE factor (.32 vs .36) but does not change the pattern for GTM4 (.17 on GTM vs .3 on ISE). Their loadings on the g_{s-1} remain relatively the same (GTM2: .7; GTM4: .75).

Table 11
Item-Factor Loadings for the Bifactor S-1 ESEM Model

	G _{S-1}	MOT	ISE	GTM	COG	SE
MOT1	0.69	0.33	0.02	0.05	0.06	0.05
MOT2	0.60	0.35	0.02	0.16	0.14	0.08
MOT3	0.69	0.32	0.05	0.18	0.12	0.04
ISE1	0.57	0.22	0.43	0.16	0.01	0.06
ISE2	0.41	0.05	0.31	0.20	-0.05	-0.02
ISE3	0.53	-0.06	0.55	-0.01	-0.04	0.11
ISE4	0.64	-0.09	0.57	-0.02	-0.10	0.16
GTM1	0.74	0.21	0.09	0.32	0.03	0.05
GTM2	0.72	0.09	0.32	0.27	-0.15	-0.03
GTM3	0.65	0.18	-0.18	0.32	0.02	-0.12
GTM4	0.75	-0.10	0.31	0.18	-0.18	-0.03
COG1	0.67	0.01	0.04	0.05	0.59	0.02
COG2	0.65	-0.02	0.03	0.01	0.64	0.01
COG3	0.60	0.07	-0.32	0.11	0.38	0.04
COG4	0.70	0.07	-0.26	0.05	0.40	-0.01
COG5	0.64	0.05	0.23	-0.22	0.36	0.05
COG6	0.67	0.11	0.01	-0.17	0.40	-0.03
SE1	0.58	0.04	0.11	-0.02	0.12	0.44
SE2	0.36	0.05	0.10	0.03	0.05	0.44
SE3	0.50	0.09	0.07	0.02	0.12	0.43
SE4	0.59	-0.01	0.06	-0.03	-0.05	0.65
SE5	0.56	-0.01	-0.12	0.01	0.01	0.49
SE6	0.61	-0.02	0.12	-0.07	-0.12	0.56
MET1	0.83	0.02	-0.08	-0.13	-0.08	0.01
MET2	0.84	0.08	-0.19	-0.23	0.03	-0.03
MET3	0.86	0.07	-0.11	-0.24	-0.01	-0.05
MET4	0.84	-0.07	-0.02	-0.13	-0.01	-0.03
MET5	0.77	-0.21	-0.14	0.09	0.01	-0.07
MET6	0.85	-0.22	-0.07	0.11	0.03	-0.06
MET7	0.77	-0.23	0.01	0.22	-0.02	-0.08
MET8	0.82	-0.16	0.03	0.06	0.10	0.00

Note. The highlighted loadings indicate the primary loadings of items in their target factor. The bold estimates were statistically significant.

Figure 5

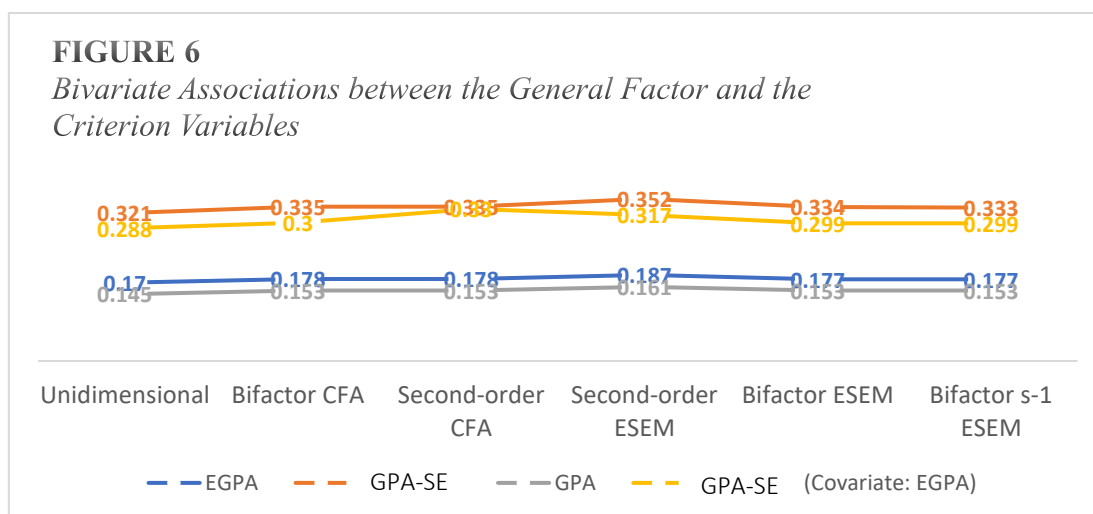
Overall, the bifactor S-1 ESEM model exhibited similar characteristic as the symmetrical bifactor ESEM but solved the anomalies in metacognitive challenges observed in the symmetrical model. Beyond this evidence, the methodological reasons for preferring the S-1 model over the symmetrical model and the theoretical evidence supporting this preference for the academic challenges scale will be further explored in the discussion section.

Criterion Validity

To explore how different types of measurement models influence the relationships between academic challenge factors and criterion variables (Expected GPA, GPA Self-efficacy, and End of Semester GPA), these associations were examined across all measurement models. Both bivariate and multivariate associations (i.e., where all facets of a measurement model

predict an outcome) were tested to understand the additive/suppressive effects of the factors with respects to each other. The parameter estimates are provided in the supplementary materials².

The bivariate associations of the general factor and the criterion variables were relatively consistent from the most basic unidimensional measurement model to the best fitting model (Figure 6). Additionally, all bivariate and multivariate associations of g factor and criterion variables were statistically significant ($p < .001$). Finally, the pattern of bivariate and multivariate associations of the g factor and the criterion variables were similar in symmetrical bifactor ESEM and the S-1 bifactor ESEM models.



Note. All associations depicted are statistically significant ($p < .001$). While the actual associations are negative, the figure displays the absolute values for ease of reading.

Abbreviations: EGPA: Expected GPA; GPA-SE: GPA Self-Efficacy; GPA: End of Semester GPA

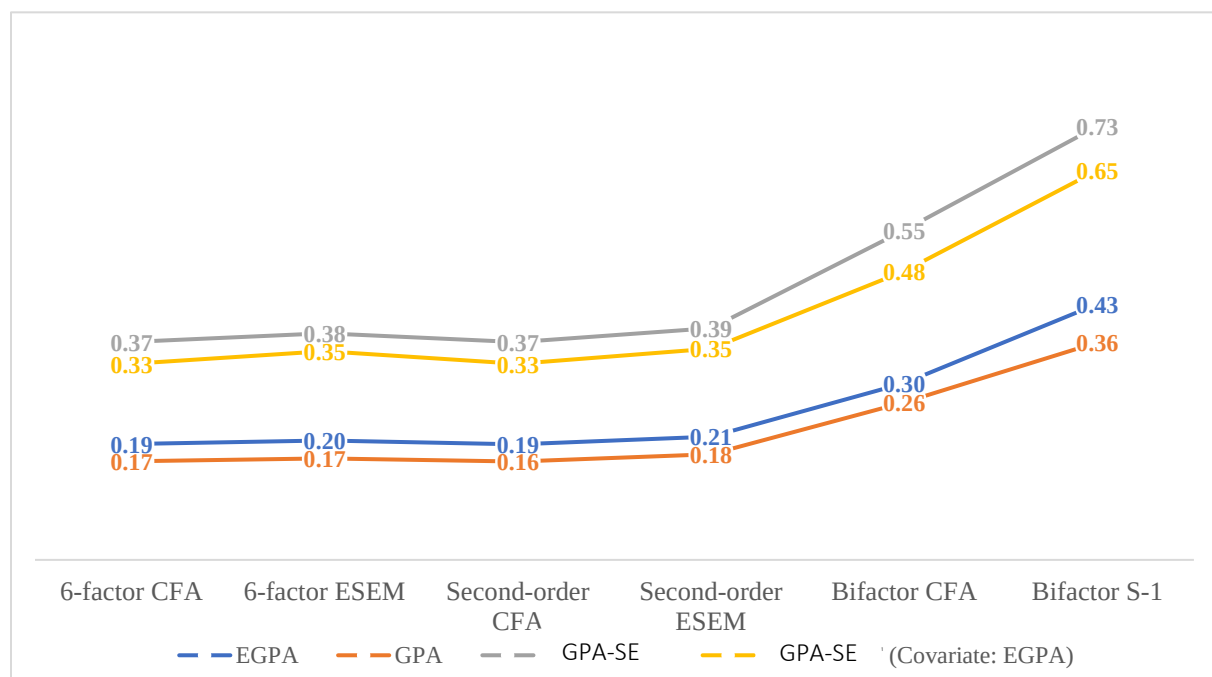
Average bivariate associations of first-order factors depicted in Figure 7 showed that the associations of the *S* factors and the criterion variables improved when the global factor is modeled as a bifactor. The highest bivariate associations between criterion variables and specific

² As discussed in the analytic strategy section, the ESEM within CFA approach was employed for conducting structural analyses described in this section.

factors were observed in the bifactor S-1 ESEM model. Additionally, the S factors exhibited noticeably higher bivariate associations with the criterion variables than the g_{s-1} factor.

Figure 7

Average Bivariate Associations between Specific Factors and the Criterion Variables



Note. All associations are statistically significant ($p < .001$). While the actual associations are negative, the figure displays the absolute values for ease of reading.

Abbreviations: EGPA: Expected GPA; GPA-SE: GPA Self-Efficacy; GPA: End of Semester GPA

Examining multivariate associations showed that S factors had limited capability in predicting criterion outcomes in the first-order models (i.e., CFA, ESEM). Similarly, S factors had limited predictive ability above and beyond that of g . While all the bivariate associations were statistically significant, when all factors concurrently predicted criterion variables, many of the associations between S factors and criterion variables were suppressed and/or became statistically non-significant (see supplementary materials). Some subscales such as GTM showed incremental validity beyond g_{s-1} in predicting the criterion outcomes. However, given the low

reliability of the specific factors in the bifactor models, these findings should be interpreted with caution.

Reliability and Dimensionality of the Bifactor S-1 ESEM Model

The reliability estimates for the bifactor S-1 ESEM model are provided in Table 2. The reliability estimates for the other measurement models are provided in the supplementary materials.

The extracted common variance showed that about 70% of common variance in all items is attributable to the general factor. This general factor demonstrated strong reliability as a unidimensional factor (ω), strong reliability when the effect of the specific factors is accounted for (ω_H), construct replicability (H) and factor determinacy (FD). All specific factors had acceptable reliability as discrete factors ($\omega = .81$ to $.93$). However, when the effect of the general factor on item variances is accounted for, the reliability of the specific factors (ω_S) became suboptimal. Particularly, ω_S for MOT and GTM was very low.

Approximately 74%-75% of the common variance in MOT and GTM (ECV_GS) was attributed to the general factor. Around 60% of the common variance in COG was due to the general factor, and about half of the common variance in SE and ISE was explained by the general factor.

Metacognitive items exhibited high I-ECV values indicating that on average about 91% of their common variance was explained by the general factor. A comparison of the reliability of the bifactor S-1 ESEM model and the symmetrical model was conducted to better understand the consequences of omitting the metacognitive factor and allowing its items to exclusively load on the general factor.

ω_S for the metacognitive challenges scale in symmetrical bifactor ESEM model was zero indicating that this scale is unreliable because its variance was almost completely explained by the g factor in the symmetrical bifactor ESEM model. To elaborate, within the symmetrical bifactor model, the ECV_SS for the metacognitive scale was .06 indicating that only 6% of common variance in the 8 items of metacognitive items can be explained by this factor while the ECV_GS was .91 indicating that 91% of its common variance was explained by the general factor. The ECV_SG was 0.2 indicating that only 2 percent of common variance in all the items of the measure is explained by the items of metacognitive subscales. All reliability indices suggested that excluding the metacognitive factor from the specific factors in the S-1 bifactor ESEM model did not negatively affected the scale's reliability; in fact, it did enhance it.

Table 11

Ancillary Reliability and Dimensionality Estimates for the Bifactor S-1 ESEM Model

	ECV_SS	ECV_SG	ECV_GS	IECV Range	Omega	OmegaH	PVR	H	FD
G	0.70	0.70	0.70	[0.39, 0.97]	0.98 ¹	0.92 ³	94.6	0.97	0.99
MOT	0.19	0.03	0.75	[0.67, 0.8]	0.81 ²	0.15 ⁴	19.0	0.27	0.39
ISE	0.40	0.07	0.52	[0.47, 0.55]	0.83 ²	0.34 ⁴	41.6	0.55	0.62
GTM	0.11	0.04	0.74	[0.7, 0.78]	0.89 ²	0.10 ⁴	11.8	0.25	0.37
COG	0.31	0.07	0.60	[0.51, 0.69]	0.93 ²	0.30 ⁴	32.7	0.66	0.67
SE	0.45	0.08	0.51	[0.39, 0.61]	0.88 ²	0.41 ⁴	46.5	0.69	0.67

Note. I-ECV range for the MET Items [.85, .97].

¹Total Omega (ω)

²Subscale Omega (ω_S)

³ Hierarchical Omega (ω_H)

⁴ Hierarchical Omega for Subscale (ω_{HS})

The average relative parameter bias (ARPB) for the S-1 bifactor model was .05, and the percentage of uncontaminated correlations (PUC) was .84³. As discussed in the analytic strategy section, this suggests that applying a unidimensional model to the multidimensional structure of the academic challenges data would result in negligible bias. In other words, the degree of

³ Since the literature on how to compute these indices for bifactor ESEM models is still developing, the Bifactor S-1 CFA was used to calculate ARPB and PUB indices.

multidimensionality present is not strong enough to undermine the interpretation of the instrument as primarily unidimensional.

Measurement Invariance of the Bifactor S-1 ESEM Model

To test the measurement invariance of the scale across sex at birth, year of study, and the two participating cohorts of the study, Multigroup Structural Equation Modeling was used⁴.

Fit improvement was initially evaluated using the DIFFTEST function of Mplus ($MD\Delta\chi^2$; Asparouhov et al., 2006). The p values for $MD\Delta\chi^2$ tests were mostly significant, favoring the fact that there was systematic difference across the groups. However, $MD\Delta\chi^2$ tends to be oversensitive to sample size. Therefore, additional indices such as change in CFI and RMSEA was used to complement chi-square difference tests (Chen, 2007; Cheung & Rensvold, 2002).

The CFI and RMSEA differences between sequentially nested model were below the ($\Delta CFI \leq 0.010$, $\Delta RMSEA \leq 0.015$) values proposed by Chen (2007) in all the three measurement invariance analyses indicating full invariance across sex at birth, year of study, and the two cohorts of the study. Additionally, the most constrained models showed strong CFI, TLI, and RMSEA indices, further supporting the invariance between groups. The invariance of measurement in terms of latent variance-covariance and latent means supported the feasibility of making valid comparisons of systematic differences between groups. The fit indices and model comparison estimates for the three multigroup analyses are reported in Table 12, Table 13, and Table 14.

⁴ Differential item functioning and differential test functioning analyses were conducted in Study 2, demonstrating invariance across the five participating universities in the sample.

Measurement Invariance Across Sex at Birth

Table 12

Measurement Invariance Across Sex at Birth (Male, N=969; Female, N=1698)

Model	χ^2 (df)	<i>p</i>	CFI	TLI	RMSEA 90%CI	SRMR	Δ CFI	Δ TLI	Δ RMSEA	Δ SRMR
Bifactor s_{-1} model for the whole sample	4329.494 (294)	0	0.967	0.947	0.072 [0.07, 0.074]	0.02				
Bifactor s_{-1} model for the Male Participants	1471.821 (294)	0	0.974	0.959	0.064 [0.061, 0.068]	0.021				
Bifactor s_{-1} model for the Female Participants	3015.526 (294)	0	0.963	0.942	0.074 [0.071, 0.076]	0.021				
Configural invariance	4411.882 (588)	0	0.968	0.95	0.07 [0.068, 0.072]	0.021				
Weak (λ) invariance	3282.211 (738)	0	0.979	0.973	0.051 [0.049, 0.053]	0.023	0.011	0.023	-0.019	0.002
Strong (λ , ν) invariance	3265.28 (825)	0	0.98	0.977	0.047 [0.045, 0.049]	0.024	0.001	0.004	-0.004	0.001
Strict (λ , ν , δ) invariance	2821.709 (856)	0	0.984	0.982	0.041 [0.040, 0.043]	0.025	0.004	0.005	-0.006	0.001
Latent variance- covariance (λ , ν , δ , ξ/ϕ) invariance	1325.34 (877)	0	0.996	0.996	0.02 [0.017, 0.022]	0.028	0.012	0.014	-0.021	0.003
Latent mean (λ , ν , δ , ξ/ϕ , η) invariance	1535.397 (883)	0	0.995	0.994	0.024 [0.022, 0.025]	0.028	-0.001	-0.002	0.004	0

Note. Chi-Square Tests for Difference Testing are reported in the supplementary materials.

Measurement Invariance Across Cohorts of study

Table 13

Measurement Invariance Across Cohorts of study (Fall intake, N=1246; Winter intake, N=1427)

Model	χ^2 (df)	p	CFI	TLI	RMSEA 90%CI	SRMR	Δ CFI	Δ TLI	Δ RMSEA	Δ SRMR
Bifactor s_{-1} model for the whole sample	4329.494 (294)	0	0.967	0.947	0.072 [0.07, 0.074]	0.02				
Bifactor s_{-1} model for the Cohort 1	2133.364 (294)	0	0.969	0.95	0.071 [0.068, 0.074]	0.021				
Bifactor s_{-1} model for the Cohort2	2233.847 (294)	0	0.969	0.951	0.068 [0.065, 0.071]	0.02				
Configural invariance	4363.398 (588)	0	0.969	0.951	0.069 [0.067, 0.071]	0.021				
Weak (λ) invariance	3225.806 (738)	0	0.98	0.974	0.05 [0.048, 0.052]	0.022	0.011	0.023	-0.019	0.001
Strong (λ , v) invariance	3102.105 (825)	0	0.981	0.979	0.045 [0.044, 0.047]	0.023	0.001	0.005	-0.005	0.001
Strict (λ , v , δ) invariance	2456.475 (856)	0	0.987	0.986	0.037 [0.036, 0.039]	0.024	0.006	0.007	-0.008	0.001
Latent variance- covariance (λ , v , δ , ξ/φ) invariance	1244.474 (877)	0	0.997	0.997	0.018 [0.015, 0.020]	0.026	0.01	0.011	-0.019	0.002
Latent mean (λ , v , δ , ξ/φ , η) invariance	1270.083 (883)	0	0.997	0.997	0.018 [0.016, 0.020]	0.026	0	0	0	0

Note. Chi-Square Tests for Difference Testing are reported in the supplementary materials.

Measurement Invariance Across Year of Study

Table 14

Measurement Invariance Across Year of Study (1st year, N=1536; 2nd year, N=1137)

Model	χ^2 (df)	p	CFI	TLI	RMSEA 90%CI	SRMR	Δ CFI	Δ TLI	Δ RMSEA	Δ SRMR
Bifactor S-1 model for the whole sample	4329.494 (294)	0	0.967	0.947	0.072 [0.07, 0.074]	0.02				
Bifactor S-1 model for the 1st year participants	2614.449 (294)	0	0.967	0.947	0.072 [0.069, 0.074]	0.021				
Bifactor S-1 model for the 2nd year participants	1846.897 (294)	0	0.97	0.952	0.068 [0.065, 0.071]	0.02				
Configural invariance	4452.534 (588)	0	0.968	0.949	0.07 [0.068, 0.072]	0.021				
Weak (λ) invariance	3286.828 (738)	0	0.979	0.973	0.051 [0.049, 0.053]	0.023	0.011	0.024	-0.019	0.002
Strong (λ , ν) invariance	3024.054 (825)	0	0.982	0.979	0.045 [0.043, 0.046]	0.023	0.003	0.006	-0.006	0
Strict (λ , ν , δ) invariance	2362.41 (856)	0	0.987	0.986	0.036 [0.035, 0.038]	0.024	0.005	0.007	-0.009	0.001
Latent variance- covariance (λ , ν , δ , ξ/ϕ) invariance	1208.278 (877)	0	0.997	0.997	0.017 [0.014, 0.019]	0.025	0.01	0.011	-0.019	0.001
Latent mean (λ , ν , δ , ξ/ϕ , η) invariance	1210.543 (883)	0	0.997	0.997	0.017 [0.014, 0.019]	0.025	0	0	0	0

Note. Chi-Square Tests for Difference Testing are reported in the supplementary materials.

Discussion

In the present study, we sought to extend prior validation of the Academic Challenges Scale (Hadwin et al., 2024) as a correlated factor model. The substantial correlations observed among the six factors of the scale suggested convergence, indicating that these factors do indeed measure aspects of the same underlying construct. However, this high degree of correlation also hinted at the presence of an overarching factor influencing responses across these dimensions. Neglecting to account for such a general factor could undermine the specificity of the scale and lead to biased parameter estimates (Asparouhov & Muthén, 2009).

To examine the possibility of considering a general construct of academic challenges, we explored both higher-order and bifactor measurement models. The bifactor model proved superior because it enabled the properly disaggregated coexistence of both general and specific factors, facilitating a more nuanced dissection of these constructs.

Another critical limitation of the measurement model for the academic challenges scale that this study sought to relax relates to the conventional assumption in CFA that factors operate as independent clusters. For instance, in a CFA model, an item such as "struggling with believing I can do my work" is attributed solely to the motivation challenges scale. However, such an item could also reflect non-primary but significant influences from other factors, such as metacognitive challenges (e.g., low confidence because they do not know how to tackle the task strategically). Such potential interdependency suggests that items may not exclusively belong to a single factor but rather exhibit significant cross-loadings. Forcing the association of items with only their primary factors in a standard CFA model could result in inflated parameter estimates, potentially distorting subsequent structural path estimates (Mai et al., 2018; Marsh et al., 2014; Morin et al., 2020; Wei et al., 2022; Zhang et al., 2023).

Bifactor and second-order models address the first issue and the ESEM approach was employed to allow for estimation of cross loadings in a confirmatory measurement model. By employing a sequential approach (Morin, 2023; Morin et al., 2016; Morin et al., 2020), we addressed these limitations both individually and collectively.

In this study, after a thorough evaluation as detailed in the analytic strategy section, the bifactor exploratory structural equation modeling (ESEM) emerged as the best-fitting model. This determination was based on global fit indicators and comparative assessments of local parameter estimates across various models, including first-order and bifactor models, as well as CFA and ESEM models. In a bifactor ESEM model, both cross-loadings and a general factor are modeled leading to a theoretically more satisfying representation of the Academic Challenges Scale. This approach lessens biases inherent in previous analyses.

Within the selected bifactor ESEM framework, an important finding was the strong loading of metacognitive challenges items on the general factor, accompanied by anomalous loadings on the specific factor of metacognitive challenges itself. This phenomenon, often referred to as the vanishing of specific factors (Morin et al., 2020; van Zyl & ten Klooster, 2021), occurs when the general factor accounts for a significant portion of the common variance among items. Consequently, there remains minimal variance for the specific target factor to explain. This results in minimal and at times anomalous patterns of factor loadings on the specific factor.

Consequently, given the dominance of the general factor in capturing the variance attributed to metacognitive challenge items, the specific factor corresponding to metacognitive challenges was omitted. This modification led to all eight indicators of this scale being treated as primary indicators of the general factor, adopting a modeling approach known as the S-1 bifactor model. In this framework, rather than interpreting the covariance among observations as a

general factor *post hoc*, a referent domain is selected *a priori*, and the general factor (g_{s-1} factor) is anchored to this scale. In other words, choosing the reference factor *a priori* in bifactor S-1 models leads to a more interpretable and replicable G-factor, compared to determining *post hoc* what the shared variance represents in symmetric bifactor models (Eid et al., 2017; Eid, 2020).

According to the stochastic measurement theory, for reflective measurement models to be valid, the indicators/domains that the latent factor is assumed to cause should be interchangeable. Factors are deemed interchangeable if they are random (not fixed) indicators of the general factor and are therefore functionally interchangeable. Randomness here means once a researcher develops a comprehensive list of items assumed to be symptoms (reflectors) of a latent factor, different sets randomly selected from that list should result in a substantially similar latent structure. Within a bifactor model, this involves assessing whether various specific factors that reflect a broader construct can be considered functionally equivalent. Functional equivalence is related to interchangeability. It means that there should be no theoretical causal relationship between indicators/domains of a construct in a measurement model because the only causal relationships allowed are between latent factors and indicators, not between latent factors or between indicators. This problem is amplified when a whole factor can be plausibly seen as being caused by other factors. Consequently, each factor in a bifactor model should contribute roughly equally to the overall g , and the g factor should be understood without needing to functionally distinguish between its constituent domains (Bollen & Bauldry, 2011; Bollen & Lennox, 1991; Eid et al., 2017; Eid, 2020; Jarvis et al., 2003).

In a symmetrical bifactor model, the G-factor represents variance shared among all indicators, while S-factors capture variance shared among all indicators of a subscale beyond what is explained by the G-factor (Morin, 2021). When specific factors differentially contribute

to the general factor and they are not functionally equivalent, the meaning of the G-factor can become ambiguous. Therefore, the actual meaning of the G-factor in the symmetrical model must be inferred on a study-by-study basis by examining the factor loadings on the G-factor. Additionally, the variance shared among all indicators as an overarching G-factor may include other sources of irrelevant common variance (e.g., common method bias).

In an S-1 model where the referent factor is chosen based on a solid and valid theoretical rationale, factors exhibit clearer conceptual boundaries and stability, as the meaning of the G-factor is primarily and directly defined by the items of the reference factor, making it less likely to fluctuate from one study to another compared to when the G is captured in a data-driven way in symmetrical bifactor models (Haywood et al., 2021; Heinrich et al., 2020; 2023). The meaning and interpretation of G-factor in an S-1 model do not change when other specific factors are added or removed from the model, as long as the reference domain remains consistent. This characteristic can facilitate more efficient knowledge accumulation in systematic reviews and meta-analyses.

Similarly, the interpretation of S-factors is clearer in an S-1 model than in symmetrical bifactor models. S-factors in S-1 models represent the part of the specific factor that cannot be explained by the reference factor—not the part of the specific factor that cannot be explained by a representation of the shared variance (G-factor) whose meaning is to be interpreted *post hoc*. As previously explained, the correlations between S-factors in symmetrical bifactor models must be fixed to zero to allow for disaggregation of G and S factors. In contrast, correlations between factors can be estimated and interpreted in bifactor S-1 models. A correlation between two specific factors represent what two S-factors have in common once the effect of the reference factor is accounted for (Bureau et al., 2023; Eid et al., 2017). For example, in our study, the

motivation scale within the bifactor S-1 ESEM model specifically measures motivational challenges independent of metacognitive influences.

A practical advantage of the bifactor S-1 compared to symmetrical bifactor models is avoiding anomalous results in empirical applications (e.g., vanishing factors). A bifactor S-1 can be defined in such a way that it is a reformulation of a first-order correlated factor model with the same levels of global fit. Because first-order factor models usually do not show anomalous results, the bifactor (S-1) model is typically also not affected by the estimation problems (Eid et al., 2018).

Given that the meaning of the general and specific factors remains fairly consistent across similar studies as long as the reference factor remains consistent, it is important to note that the reference factor should be chosen based on theory and not statistical evidence such as model fit or parameter estimates (Burns et al., 2020; Eid, 2020; Eid et al., 2017; Heinrich et al., 2020). To elaborate, the selection of a referent domain should be guided by theoretical considerations and not solely employed as a methodological remedy for anomalies observed in a symmetrical bifactor model. In our case, the choice of metacognitive challenges as the referent domain was originally a response to the failure of the symmetrical model. Theoretical underpinnings from self-regulated learning (SRL) suggest that SRL is driven predominantly by metacognitive processes, with metacognitive monitoring and control positioned as central components (Panadero, 2017; Winne, 1996; 2021; Winne & Hadwin, 1998). Closer examination of the metacognitive challenge items showed that these challenges could be seen as representing higher-order regulatory challenges that likely originate from other types of academic challenges.

This understanding informs the interpretation of psychometric redundancy observed between the general factor and the metacognitive factor in the symmetrical bifactor model,

suggesting that like the general factor, metacognitive challenges may encapsulate higher-order challenges compared to other specific factors within the scale. Similarly, the analysis of composite reliability for the bifactor S-1 ESEM model, compared with that of the symmetrical bifactor ESEM model, underscored the psychometric redundancy between the general factor and the metacognitive challenges factor (MET), further supporting the decision to retain the S-1 model. Finding from the unidimensional polytomous IRT validation of the scale (Study 2) also supported that an overall total score for the scale predominantly derives its meaning from the metacognitive items. However, this decision was not based on a solid *a priori* theory and should be interpreted with caution. The limitations of this decision, along with potential ways to address them in future versions of the scale, are detailed in the chapter 4 of the dissertation under the title “Prioritizing Ecological Validity in Defining the General Factor”.

Moreover, variance decomposition indices across the general and specific factors revealed that approximately 70% of the common variance in the observed items is attributed to the general_{S-1} factor. This substantial contribution by the general factor highlights its dominance in explaining the overall variability in the scale's items. In contrast, the specific factors exhibited suboptimal reliability when considered above and beyond the influence of the general factor on item variances. This observation suggests that while the scale does embody multidimensionality, the presence of a strong general factor adequately captures the essence of the academic challenges being assessed. Overall, these ancillary analyses of reliability and dimensionality (Dueber, 2017; Rodriguez et al., 2016) revealed that despite the inherent multidimensionality of the scale, the existence of a robust general factor allows for modeling the scale as essentially unidimensional without significantly compromising the accuracy of the scale in measuring academic challenges.

The analysis of bivariate associations between the factors and criterion variables across different measurement models indicated that a representation of the shared variance among observations as a general factor demonstrated a consistent associations with criterion variables, irrespective of the measurement modeling approach used. This consistency across models attests to the robustness of the general factor, suggesting that it effectively captures a core aspect of academic challenges that is predictively relevant across various analytical methods.

Further investigation into the bivariate associations of the specific factors revealed that these factors are more strongly associated with the criterion variables in bifactor models where they represent a variance uncontaminated by that of the general factor. Notably, the specific factors exhibited the highest bivariate associations with the criterion variables in the bifactor S-1 ESEM model. The magnitude of these associations was higher than those observed for the general factor with the same criterion variables, suggesting that, despite their suboptimal reliability, the variance in the specific factors uncontaminated by the general factor represents meaningful constructs rather than merely method effects that can occur in the bifactor models.

Additionally, the multivariate analyses which explored associations among the specific factors themselves and between the specific and general factors revealed that once the effect of the general factor was accounted for, the specific factors no longer showed major significant associations with the criterion variables. This suppression of associations (i.e., effect sizes) between specific factors and criterion construct and the nonsignificant *p*-values when the general factor is introduced to the regression equations suggest that while the specific factors were found to be meaningfully linked to the criterion variables as independent constructs, they do not provide incremental validity beyond the general factor in explaining these relationships.

Although the specific factors did not contribute additional explanatory power beyond the general factor when they all concurrently predicted the criterion variables along with the general factor, their significant bivariate associations with the criterion variables support their potential utility as independent correlates of the criterion constructs. This finding is in line with a previous study that have validated these specific factors as independent constructs using a graded response polytomous IRT model (Rostampour et al., in preparation).

Finally, supporting evidence of measurement invariance demonstrated that the bifactor S-1 ESEM model function similarly, irrespective of the respondent's sex at birth, their academic year of study, or the cohort to which they belong. The invariance of measurement in terms of latent variance-covariance and latent means supported the feasibility of making valid comparisons of systematic differences between groups.

Table 15 provides a summary of findings and Table 16 provides a summary of the recommendations for using the scale for practical and research purposes.

Table 15
Summary of the Findings

Objective	Strategies tested	Strategies Chosen	Reason(s)
Modeling a General Factor	Second-order and Bifactor Measurement Models	Bifactor	• Higher Fit • Properly disaggregation of S and G factors
Modeling Cross Loadings	Correlated CFA and Correlated ESEM compared	Correlated ESEM	• Higher Fit • Improved ecological validity
Modeling a General Factor + Modeling Cross-loadings	Bifactor ESEM, Bifactor S-1 ESEM	Bifactor S-1 ESEM	• Improved ecological validity • Theoretical anchoring of the G factor • Higher Stability and replicability • Removing the psychometric redundancy of the G and metacognitive challenges • Less likelihood of facing statistical anomalies

Objectives	Findings
Composite Reliability and Dimensionality of the S-1 Bifactor ESEM Model	• approximately 70% of the common variance in the observed items is attributed to the general_{S-1} factor • The scale can be operationalized as essentially unidimensional • Suboptimal reliability for some specific factors (e.g., GTM)
Bivariate associations between the g and the criterion variables across measurement models	• Various representation of the g demonstrated a consistent association with criterion variables, irrespective of the modeling approach used
Bivariate associations between the S factors and the criterion variables across measurement models	• the specific factors exhibited the highest associations with the criterion variables in the bifactor S-1 ESEM model • The magnitude of these associations was higher than those observed for the general factor with the same criterion variables
Multivariate analyses exploring associations between the specific factors and the criterion variables	• Despite strong and significant bivariate associations with the criterion variables, specific factors had limited predicted power in predicting criterion outcome when these variables were regressed on all specific factors at once.
Multivariate analyses exploring associations of the specific factors and the general factor with the criterion variables	• Specific factors did not provide much incremental validity beyond the general factor in explaining these relationships. • GTM demonstrated incremental validity in predicting criterion validity above and beyond g.
Measurement invariance	• the bifactor S-1 ESEM model function similarly, irrespective of the respondent's sex at birth, their academic year of study, or the cohort to which they belong. • Mean differences can be compared for these demographic groups

Table 16*A summary of How to Use the Academic Challenges Scale*

Research Question demands operationalization of...	Is SEM Available?	Recommendation
A Single Specific Challenge	Yes	Model it as a unidimensional construct
A Subset of specific challenges	Yes	Model them as unidimensional constructs that are allowed to covary (ideally using Correlated ESEM)
All specific challenges And/or General Academic Challenges	Yes	Model as S-1 bifactor ESEM with metacognitive challenges as a reference domain (note the low reliability of some subscales in the current version of the scale)
Specific Challenges	No	Model as average subscale scores
Overall Academic Challenges	No	Model as an average total score
Shorter representation of the overall score needed	Yes/No	Use 8 items of the metacognitive challenges scale for a shorter but highly correlated representation of the overall score

Suggestions For Future Studies

The study's findings provide several insights that suggest directions for both advancing research on academic challenges and improving the academic challenges scale. Based on the observed dynamics between general and specific factors in bifactor models, I recommend specific methodological adjustments and extensions to improve the scale's utility and applicability.

Developing Discrete Measures for Specific Challenges. The results highlighted that while both general and specific factors are valid and reliably associated with criterion variables, the specific factors exhibit reduced reliability within the bifactor models. This reduction is due to the

high interconnectedness among the challenges and the dominance of the general factor that results in the general factor absorbing much of the variance that might otherwise be attributed to specific factors.

To address this issue, it is suggested that specific subscales be extended as discrete measures of academic challenges. By designing scales that include more detailed indicators specific to each challenge area, the unique variance in these specific scales can be boosted, potentially making both general and specific factors to be reliably represented in the bifactor models. Additionally, while empirical proof will be needed, such enhanced specificity could make the scale a more effective independent measure of challenges and may enhance its practical utility and diagnostic capabilities.

Refining Bifactor Measurement Model for the Scale. According to stochastic measurement theory, in a reflective measurement model—where latent factors are presumed to cause the indicators—the indicators should ideally be interchangeable (Eid et al., 2017). This suggests a certain level of functional equivalence in how they manifest the underlying latent construct. Non-interchangeable indicators, on the other hand, might be more aptly modeled using a variant of the formative linear combination approach, where indicators form the latent factor rather than reflect it (Hair et al., 2017).

In the context of the study, the selection of the metacognitive scale as a referent factor in the bifactor S-1 model was done *post hoc* though it was aligned with the empirical support from Item Response Theory validation of the scale (Rostampour et al, In preparation) as well as theoretical considerations. As discussed, both empirical and theoretical evidence supports that the metacognitive challenges factor represents a higher-level functional layer of challenges compared to other factors.

Similarly, some items from the Goal and Time Management (GTM) factor, such as GTM2 and GTM4, operate at a different granularity compared to other items within the scale. For example, items like "I struggled with managing time and task" are at a granularity level that might better serve as indicators of general academic challenges rather than a specific factor. In other words, the failure to manage tasks appears to be a better reflection of a pervasive failure across various regulatory domains, rather than reflecting a dysregulation within a single specific regulatory domain.

To enhance the GTM scale's alignment with the broader framework of academic challenges, modifications could involve incorporating items that reflect challenges related to various aspects of behavioral regulation, as identified in Rostampour et al. (in preparation) study. This adjustment would allow the indicators such as GTM2 and GTM4 to be exclusively explained by the general factor (g), enhancing the model's specificity and clarity.

Alternative Measurement Approach for S-1 Bifactor model. If a symmetrical bifactor model is considered preferable, the inclusion of items that explicitly measure general academic challenges—such as "...during the past two weeks, I generally found my academic work challenging"—could anchor the general factor effectively. This approach would allow for a clearer distinction between general and specific challenges without the need to collapse any of the specific scales (Eid et al., 2017). Such a model might exhibit fewer anomalies compared to the traditional symmetrical bifactor model because the general factor is defined *a priori* rather than interpreted *post hoc*.

Interpreting Beliefs and Actions in the Context of Perceived Challenges. Academic challenges belong to a self-diagnostic tool called SRL-PSD that integrate scales that measure what students do (self-regulated learning practices), what they believe (importance of SRL and

SRL self-efficacy), and the challenges they perceive (Hadwin et al., 2024). Future research should explore how these aspects are associated with each other both cross-sectionally and longitudinally. Situating SRL practices and beliefs within the context of the challenges students experience can offer deeper insights into how students intentionally navigate their academic environments. Understanding the interaction between contextual factors and SRL components in relation to academic challenges can help identify more specific areas for intervention that are tailored to diverse student needs.

Within-Person Change. The current study examined the composite reliability of the academic challenges scale. However, to further enhance the utility of the scale, particularly in educational settings where both the personal (e.g., goals, standards) and contextual (e.g., learning demand) can alter over time, it is crucial to evaluate within-person change, or test-retest reliability.

Evaluating within-person change would enable researchers and educators to track how academic challenges evolve throughout different academic cycles such as over a semester or an academic year. This can reveal if certain challenges are more prevalent at specific times, such as early or late in the semester, and whether these patterns are consistent across cohorts of students. For instance, students who respond to the survey at the second week may face social and emotional difficulties, typical of early university life when adapting to new social settings is common. By the fourth week, students may perceive more cognitive and metacognitive challenges as academic demands increased. This longitudinal approach would offer a more dynamic view of how academic challenges and students' responses to them change over time, potentially influenced by mid-term assessments, feedback, and the accumulation of academic experiences. Such understanding could lead to more timely and effective support strategies for students at different stages of the academic cycle.

Repeated measurement can also shed light on the contextual factors that contribute to any possible changes in response patterns across measurement occasions. If this evaluation is conducted alongside the collection of data on potential personal (e.g., goals, standards, expectations, etc.) and contextual influences (e.g., changes in course load, personal stressors, type of the tasks, etc.), it could provide a more comprehensive understanding of what factors most significantly impact students' academic experiences.

Currently, the scale uses a two-week window to assess academic challenges, assessing test-retest reliability can help determine whether this timeframe is sufficient. It is possible that a more prolonged measurement window could capture more persistent challenges, rather than transient ones that might not have a lasting impact on a student's overall academic performance and wellbeing.

Validating the Measure Using a Mixed Method Design. Psychometric validation studies can be accompanied by qualitative interview studies within a mixed-methods design to corroborate the validation of a researcher's/diagnostician's interpretations about the scale. Interviews can be conducted after participants complete the questionnaire. During the interview, participants can be asked about their major academic challenges, and their responses can be compared with the survey results to assess alignment both in terms of the level and categorization of the challenges.

Additionally, students' responses to the interview questions can be contextualized within the items themselves. For instance, an item is first read aloud, and then participants are asked to explain why they selected a specific response or what does “mostly” mean to the participant and what are the exceptions the participants choose to discount in their memories. This approach helps understand what came to their mind when responding to the item. Participants can also provide examples to further contextualize their answers. A mix of open-ended and structured

questions could be used to elicit detailed responses and gain deeper insights into the participants' experiences and perspectives.

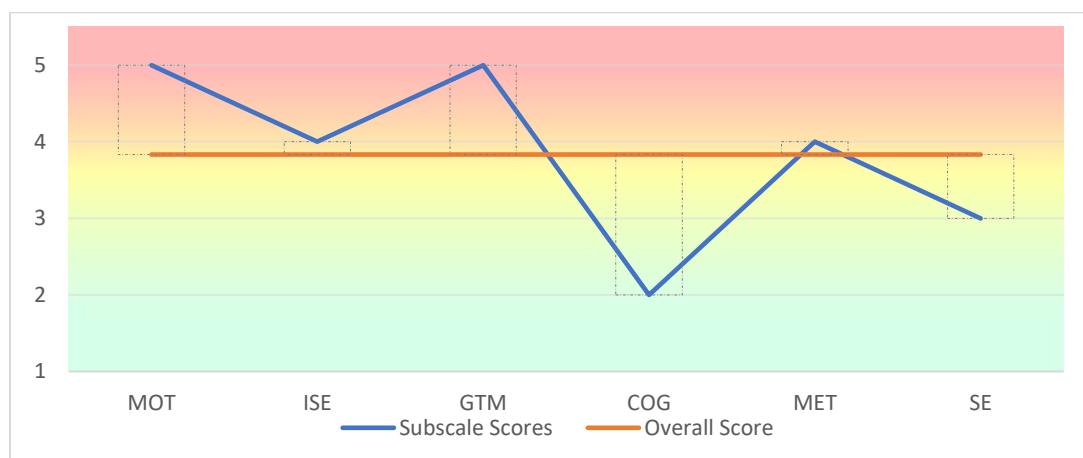
Computing Manifest Scores. Given the advanced measurement model and the practical objectives of the scale, it is crucial to determine how to compute manifest scores using simple arithmetic in practical settings where Structural Equation Modeling (SEM) is not available. These recommendations were made in Table 16; however, other opportunities also exist that need further investigation.

In the context of bifactor models, scoring and interpretation of the factors, particularly the general factor and specific factors, necessitate careful consideration due to their inherent psychometric properties and interrelationships. The g factor in a bifactor model typically captures the common variance among all items across the different domains being measured. Theoretically, it could be computed as a sum or average of all items, ideally using a weighted approach where weights reflect the loadings of items on the g factor. This method ensures that items that more strongly represent the common variance across domains have a greater influence on the calculated score.

Scoring for S factors, which represent the variance specific to each domain after accounting for the g factor, can similarly be calculated as averages of scores on each specific factor. However, interpreting these scores requires additional considerations. Scores on S factors should be viewed not only as standalone indicators of domain-specific challenges but also in relation to the g factor. For example, a higher score on an S-factor relative to the g factor score indicates a positive deviation, suggesting that the specific challenge is more pronounced than what would be expected based on the general level of challenges. In practice, it might be useful to interpret deviation scores for S factors, which represent how much an individual's score on a specific factor deviates from their score on the g factor (see Figure 8).

Figure 8

How to Interpret an Individual's Subscale Scores in Relation to their Overall Score



Note. The green, yellow, and red background represent norm-based cut-offs for low, moderate, and high challenges. The dotted bars represent how much an individual's score on a specific factor deviates from their score on the g factor.

There is no straightforward way to compute non-refined composite scores in the S-1 bifactor model because of the psychometric overlap between the referent factor and the g factor. The same method described above that applies to the symmetrical bifactor works for the bifactor S-1 model as well. However, interpretation and subsequent use of scores must be done with the awareness that the referent domain's score and the g factor's score are psychometrically overlapping.

It is important to recognize that the suggested approach does not adequately disentangle the specific and general challenges. Therefore, future research is necessary to validate such non-refined composite scores for coexisting general and specific factors in relation to norm-based cut-offs, ensuring they offer meaningful and practical information.

Limitation

This study, while providing useful perspectives into the measurement of academic challenges, has several limitations that warrant consideration.

This study focused on students' expected academic performance, self-efficacy for performance, and actual academic performance (end-of-semester GPA) to study the criterion validity of the academic challenges scale. The observed small to medium effect sizes were expected, considering that challenges perceived during the first weeks of university should predict their actual academic performance at the end of the semester. Future studies could incorporate a wider array of constructs that potentially mediate/moderate the associations of academic challenges with academic outcomes, student success, and well-being.

Also, the calibration between students' expectations and their actual performance is another important variable that was not examined in this study. Understanding how well students predict their performance may provide insights into their self-awareness and metacognitive monitoring abilities. Also, future research should track how students' expectations and perceptions of academic challenges evolve throughout the semester.

The conversion of GPA into a standardized 5-point scale across different universities in the study introduces several methodological considerations and potential biases, which are crucial to acknowledge when interpreting the findings. The necessity to convert GPA scores to a uniform scale to accommodate less detailed GPA scales at some participating universities could potentially distort the relative performance levels of students, particularly if the original scales differ significantly in granularity. However, the numerical values were converted to the letter grades based on the site-specific general guidelines provided by each institution to ensure the meaning remains consistent even though the granularity changes.

The distribution of GPA data revealing a predominance of high-performing students (median=3) suggests a skewed sample which may not accurately represent the broader student population. This skew towards high performers limits the generalizability of the findings, especially when considering criterion validity for students with a broader range of academic performance. Future studies might need to stratify sampling or employ weighting methods to ensure a more representative sample. Additionally, the significant amount (33%) of missing GPA data and its uneven distribution across sites may have affected the study's outcomes and interpretations. Such missing data, arising from varying consent processes and institutional regulations, can introduce bias, especially if the missingness is systematic (e.g., lower-performing students being less likely to consent to their GPA data being collected).

The complications in calculating semester GPA due to different frequencies of year-long courses across sites further complicated the analysis. This is because, for these courses, part of the GPA awarded in the second semester is, in fact, attributable to the work completed in the previous semester. Future studies examining academic performance should consider these complexities, especially when comparing students across various academic structures and schedules. That being said, the study's findings align with those of an earlier study by Hadwin et al. (2024), which distributed the survey towards the end of an academic semester to 476 students at a single university. However, the effect sizes in the Hadwin et al. (2024) study were slightly larger.

The data for this study were collected during the COVID-19 pandemic, specifically in the Fall of 2021 and Winter of 2022. Throughout this period, all participants engaged in university courses through online, off-campus formats. For most students and instructors, this shift to online education was a new experience. The student participants were geographically spread out,

residing across Canada and internationally, predominantly from their homes. Additionally, the extent of COVID-19-related restrictions on social activities varied widely both within Canada and globally. Many extracurricular activities were either canceled or significantly restricted. These exceptional circumstances should be carefully considered when interpreting the results of the study. Further research under different conditions is recommended to validate these findings.

This study used the same dataset as a previous investigation (Rostampour et al., in preparation). This decision should not be interpreted as data overuse, as the two studies are complementary in nature. The initial analysis focused on a detailed item-level examination using graded response Item Response Theory (IRT) to assess measurement precision. In contrast, the current study explored the multidimensionality of the construct. These distinct focuses—precision and structure—provide a comprehensive understanding of the scale by maximizing insights through different analytical lenses.

Moreover, the full measurement invariance across the two cohorts (Fall intake N=1246, Winter intake N=1427) underscores the credibility of using this dataset and the potential applicability of the findings across similar populations (See section “Limitations” of chapter 4 of dissertation for a discussion on potential shortcomings of exclusively relying on demographic variables as distinctive characteristics of populations”). Findings from the measurement invariance supported configural, metric, scalar, strict, latent variance-covariance, and latent means invariance across these two subsamples and indicated that even if the analyses were conducted with different cohorts (cohort 1 for the initial study and cohort 2 for the current study), the findings would have been consistent. Using the same sample for the two studies allowed for examining the methodological synergies of the analytical methods employed.

However, researchers must be cautious to avoid overfitting models to a specific dataset, a risk that increases with repeated data use. Overfitting can make models too tailored to one set of data, potentially reducing their relevance to other datasets or real-world scenarios. Future studies can therefore use new samples with diverse characteristics to validate the results further.

To further investigate the multidimensionality of the scale and in line with its diagnostic objectives, a complementary person-centered validation study is appended as an addendum. Interested readers are invited to review this study as well.

Ethics Statement

The study is funded through Social Sciences and Humanities Research Council (SSHRC) Partnership Grant [Grant#: 895-2019-1021]. The study is ethically approved by the five participating data collection sites by their respective ethic review boards [Dalhousie University (2021-5689), Saint Francis Xavier University (ROMEO File #25486), University of Montreal (2020-883), York University (2021-272) and University of British Columbia – Okanagan (H21-02007)]. Students' consent to access their institutional data from the registrar's office of their universities was obtained separately from their general consent to participate in the study. All non-consenting students were removed from the dataset. The dataset was fully anonymized, cleaned, and shared across sites after evaluating all ethical considerations and the existing regulations enforced by the relevant Research Ethics Boards.

References

- Abdi, H. (2010). Holm's sequential Bonferroni procedure. In N. Salkind (Ed.), *Encyclopedia of Research Design* (Vol. 1, pp. 1-8). SAGE Publications.
<https://doi.org/10.4135/9781412961288.n178>
- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 397-438.
<https://doi.org/10.1080/10705510903008204>
- Asparouhov, T., Muthén, B., & Muthén, B. O. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. *Matrix*, 1(5), 1-6.
<https://doi.org/10.1080/10705510903008204>
- Beauducel, A. (2011). Indeterminacy of factor score estimates in slightly misspecified confirmatory factor models. *Journal of Modern Applied Statistical Methods*, 10(2), 583-598. <https://doi.org/10.22237/JMASM%2F1320120900>
- Boekaerts, M. (2011). Emotions, emotion regulation, and self-regulation of learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 422-439). Routledge. <https://doi.org/10.4324/9780203839010.CH26>
- Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, 16(3), 265-284.
<https://doi.org/10.1037/a0024448>
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305-314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, 5(1), 184-186.
<https://doi.org/10.1177/2167702616657069>
- Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16(1), 129-132.
<https://doi.org/10.1080/00401706.1974.10489158>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.
- Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, 80(4), 796-846. <https://doi.org/10.1111/j.1467-6494.2011.00749.x>

- Bureau, J. S., Guay, F., Plamondon, A., Ratelle, C. F., Howard, J. L., & Gilbert, W. (2023). Empirical testing of an alternative modeling of the self-determination continuum. *Motivation and Emotion*, 47(1), 46-60. <https://doi.org/10.1007/s11031-022-09976-9>
- Burke, J. D., & Johnston, O. G. (2020). The bifactor S-1 model: A psychometrically sounder alternative to test the structure of ADHD and ODD? *Journal of Abnormal Child Psychology*, 48, 911-915. <https://doi.org/10.1007/s10802-020-00645-4>
- Butler, D. L. (2023). Qualitative approaches to investigating self-regulated learning: Contributions and challenges. In D. H. Schunk & J. A. Greene (Eds.), *Using qualitative methods to enrich understandings of self-regulated learning* (pp. 59-63). Routledge. <https://doi.org/10.1207/00461520252828564>
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245-281. <https://doi.org/10.3102/00346543065003245>
- Cambridge-Williams, T., Winsler, A., Kitsantas, A., & Bernard, E. (2013). University 100 orientation courses and living-learning communities boost academic retention and graduation via enhanced self-efficacy and self-regulated learning. *Journal of College Student Retention*, 15(2), 243-268. <https://doi.org/10.2190/CS.15.2.f>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
- Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41(2), 189-225. https://doi.org/10.1207/s15327906mbr4102_5
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools*, 41(5), 537-550. <https://doi.org/10.1002/PITS.10177>
- Cleary, T. J., & Zimmerman, B. J. (2012). A cyclical self-regulatory account of student engagement: Theoretical foundations and applications. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 237-257). Springer. https://doi.org/10.1007/978-1-4614-2018-7_11
- De Beer, L. T., & Morin, A. J. S. (2022). (B)ESEM invariance syntax generator for Mplus. Retrieved from https://stattools.app/b_esem/

- De Beer, L. T., & Van Zyl, L. E. (2019). ESEM code generator for Mplus. Retrieved from <http://www.surveymhost.co.za/esem/> doi: 10.6084/m9.figshare.8320250
- Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of Welch's F-test instead of the classical F-test in one-way ANOVA. *International Review of Social Psychology*, 32(1), 13. <https://doi.org/10.5334/IRSP.198>
- Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal*, 15(1), 292-300. <https://doi.org/10.1177/1536867X1501500117>
- Dueber, D. M. (2017). Bifactor indices calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. <https://doi.org/10.13023/edp.tool.01>
- Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399-412. <https://doi.org/10.1111/bjop.12046>
- Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 54, pp. 103-140). Elsevier Academic Press. <https://doi.org/10.1016/B978-0-12-385527-5.00004-8>
- Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S-1 model for individual clinical assessment. *Journal of Abnormal Child Psychology*, 48(7), 895-900. <https://doi.org/10.1007/s10802-020-00624-9>
- Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, 22(3), 541-562. <https://doi.org/10.1037/met0000083>
- Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence*, 6(3), 42. <https://doi.org/10.3390/jintelligence6030042>
- Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484-501. <https://doi.org/10.1177/2515245920951747>
- Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1(2), 113-125. <https://doi.org/10.3102/10769986001002113>

- Geiser, C., Eid, M., West, S. G., Lischetzke, T., & Nussbeck, F. W. (2012). A comparison of method effects in two confirmatory factor models for structurally different methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(3), 409-436.
<https://doi.org/10.1080/10705511.2012.687658>
- Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science*, 50(1), 21-43.
- Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, 55, 57-68.
<https://doi.org/10.1016/J.INTELL.2016.01.006>
- Gillet, N., Morin, A. J., Huart, I., Colombat, P., & Fouquereau, E. (2019). The forest and the trees: Investigating the globality and specificity of employees' basic need satisfaction at work. *Journal of Personality Assessment*, 101(5), 521-535.
<https://doi.org/10.1080/00223891.2019.1591426>
- Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. Routledge.
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430-450.
- Hadwin, A., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation, and shared regulation in collaborative learning environments. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 83-106). Routledge. <https://doi.org/10.4324/9781315697048.CH6>
- Hadwin, A. F., & Winne, P. H. (2012). Promoting learning skills in undergraduate students. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and mental structures* (pp. 201-227). Cambridge University Press.
- Hadwin, A. F., Davis, S. K., Bakhtiar, A., & Winne, P. H. (2019). Academic challenges as opportunities to learn to self-regulate learning. In D. Ifenthaler & Y. J. Kim (Eds.), *Problem solving for teaching and learning* (pp. 34-47). Routledge.
<https://doi.org/10.4324/9780429400902-4>
- Hadwin, A. F., Rostampour, R., & Winne, P. H. (2024). Advancing self-reports of self-regulated learning: Validating new measures to assess students' beliefs, practices, and challenges [Manuscript submitted for publication].
- Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017). Mirror, mirror on the wall: A comparative evaluation of composite-based structural equation modeling methods. *Journal of the Academy of Marketing Science*, 45, 616-632.
<https://doi.org/10.1007/s11747-017-0517-x>

- Hand, E. D., & Lonigan, C. J. (2022). Examining the relations between Preschooler's externalizing behaviors and academic performance using an S-1 bifactor model. *Research on Child and Adolescent Psychopathology*, 50(5), 577-589. <https://doi.org/10.1007/s10802-021-00861-6>
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 195-216). Scientific Software International. https://doi.org/10.1207/S15328007SEM0704_2
- Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., & Kirby, J. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, 95, 103208. <https://doi.org/10.1016/j.jbi.2019.103208>
- Haywood, D., Baughman, F. D., Mullan, B. A., & Heslop, K. R. (2021). Going "up" to move forward: S-1 bifactor models and the study of neurocognitive abilities in psychopathology. *International Journal of Environmental Research and Public Health*, 18(14), 7413. <https://doi.org/10.3390/ijerph18147413>
- Heinrich, M., Geiser, C., Zagorscak, P., Burns, G. L., Bohn, J., Becker, S. P., ... & Knaevelsrud, C. (2023). On the meaning of the "P factor" in symmetrical bifactor models of psychopathology: Recommendations for future research from the bifactor-(S- 1) perspective. *Assessment*, 30(3), 487-507. <https://doi.org/10.1177/10731911211060298>
- Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck depression inventory–II. *Assessment*, 27(7), 1429-1447. <https://doi.org/10.1177/1073191118803738>
- Howard, J. L., Gagné, M., Morin, A. J., & Forest, J. (2018). Using bifactor exploratory structural equation modeling to test for a continuum structure of motivation. *Journal of Management*, 44(7), 2638-2664. <https://doi.org/10.1177/0149206316645653>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, 30(2), 199-218. <https://doi.org/10.1086/376806>
- Jobst, L. J., Bader, M., & Moshagen, M. (2023). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods*, 28(1), 207-231. <https://doi.org/10.1037/met0000423>

- Kim, Y., Brady, A. C., & Wolters, C. A. (2020). College students' regulation of cognition, motivation, behavior, and context: Distinct or overlapping processes? *Learning and Individual Differences*, 80, 101872. <https://doi.org/10.1016/j.lindif.2020.101872>
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, 9(2), 202-220. <https://doi.org/10.1177/1094428105284919>
- Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48, 936-949. <https://doi.org/10.3758/s13428-015-0619-7>
- Mai, Y., Zhang, Z., & Wen, Z. (2018). Comparing exploratory structural equation modeling and existing approaches for multiple regression with latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(5), 737-749. <https://doi.org/10.1080/10705511.2018.1444993>
- Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275-340). Lawrence Erlbaum Associates.
- Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85-110. <https://doi.org/10.1146/annurev-clinpsy-032813-153700>
- Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 439-476. <https://doi.org/10.1080/10705510903008220>
- Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, 49(6), 1194-1218. <https://doi.org/10.1037/a0029121>
- McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press. <https://doi.org/10.2307/2669496>
- Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, 3(1), 2-20. <https://doi.org/10.3390/JINTELLIGENCE3010002>
- Morin, A. J. S. (2023). Exploratory structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 503-524). Guilford.

- Morin, A. J. S., & Asparouhov, T. (2018). Estimation of a hierarchical Exploratory Structural Equation Model (ESEM) using ESEM-within-CFA. *Montreal, QC: Substantive Methodological Synergy Research Laboratory*.
- Morin, A. J., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1), 116-139. <https://doi.org/10.1080/10705511.2014.961800>
- Morin, A. J. S., Blais, A. R., & Chénard-Poirier, L. A. (2022). Doubly latent multilevel procedures for organizational assessment and prediction. *Journal of Business and Psychology*, 37(1), 47-72. <https://doi.org/10.1007/S10869-021-09736-5>
- Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.). Information Age Publishing, Inc. <https://doi.org/10.1007/S10869-021-09736-5>
- Morin, A. J., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (pp. 1044-1073). <https://doi.org/10.1002/9781119568124.ch51>
- Moshagen, M., & Bader, M. (2023). semPower: General power analysis for structural equation models. *Behavior Research Methods*, 1-22. <https://doi.org/10.3758/s13428-023-02254-7>
- Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41(5), 407-422. <https://doi.org/10.1016/J.INTELL.2013.06.004>
- Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3), 431-462. <https://doi.org/10.1007/BF02294365>
- Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74-98. <https://doi.org/10.1016/J.EDUREV.2017.08.004>

- Panadero, E., Jonsson, A., & Strijbos, J. W. (2016). Scaffolding self-regulated learning through self-assessment and peer assessment: Guidelines for classroom implementation. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 311-326). Springer International Publishing.
https://doi.org/10.1007/978-3-319-39211-0_18
- Pino-Pasternak, D., Basilio, M., & Whitebread, D. (2014). Interventions and classroom contexts that promote self-regulated learning: Two intervention studies in United Kingdom primary classrooms. *Infancia y Aprendizaje*, 37(3), 336-368.
<https://doi.org/10.7764/PSYKHE.23.2.739>
- Prokofieva, M., Zarate, D., Parker, A., & Tharmarajah, S. (2023). Exploratory structural equation modeling: A streamlined step by step approach using the R Project software. *BMC Psychiatry*, 23, 546. <https://doi.org/10.1186/s12888-023-05028-9>
- Raykov, T., & Marcoulides, G. A. (2018). On studying common factor dominance and approximate unidimensionality in multicomponent measuring instruments with discrete items. *Educational and Psychological Measurement*, 78(3), 504-516.
<https://doi.org/10.1177/0013164416678650>
- Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent scales: An interrelationship decomposition approach. *Educational and Psychological Measurement*, 73(4), 581-600. <https://doi.org/10.1177/0013164412470451>
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696. <https://doi.org/10.1080/00273171.2012.715555>
- Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95(2), 129-140. <https://doi.org/10.1080/00223891.2012.725437>
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92(6), 544-559. <https://doi.org/10.1080/00223891.2010.496477>
- Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73(1), 5-26.
<https://doi.org/10.1177/0013164412449831>
- Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, 23(1), 51-67.
https://doi.org/10.1207/s15327906mbr2301_3

- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137-150. <https://doi.org/10.1037/met0000045>
- Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 1-15). Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9781315697048-1>
- Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research*, 53(5), 676-694. <https://doi.org/10.1080/00273171.2018.1476221>
- Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: SRMR vs. RMSEA. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 1-15. <https://doi.org/10.1080/10705511.2019.1611434>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74, 107-120. <https://doi.org/10.1007/s11336-008-9101-0>
- Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, 25(7), E34-E46. <https://doi.org/10.1002/da.20432>
- Smits, I. A., Timmerman, M. E., Barelds, D. P., & Meijer, R. R. (2014). The Dutch symptom checklist-90-revised: Is the use of the subscales justified? *European Journal of Psychological Assessment*, 31(4), 263-271. <https://doi.org/10.1027/1015-5759%2FA000233>
- Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183-206). Routledge. <https://doi.org/10.4324/9781315736013.CH9>
- Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69, 613-625. <https://doi.org/10.1007/BF02289858>
- Tuominen-Soini, H., Salmela-Aro, K., & Niemivirta, M. (2012). Achievement goal orientations and academic well-being across the transition to upper secondary education. *Learning and Individual Differences*, 22(3), 290-305. <https://doi.org/10.1016/J.LINDIF.2012.01.002>
- Van Zyl, L. E., & Ten Klooster, P. M. (2022). Exploratory structural equation modeling: Practical guidelines and tutorial with a convenient online tool for Mplus. *Frontiers in Psychiatry*, 12, 795672. <https://doi.org/10.3389/fpsy.2021.795672>

- Wei, X., Huang, J., Zhang, L., Pan, D., & Pan, J. (2022). Evaluation and comparison of SEM, ESEM, and BSEM in estimating structural models with potentially unknown cross-loadings. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(3), 327-338. <https://doi.org/10.1080/10705511.2021.2006664>
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4), 330-336. <https://doi.org/10.1093/BIOMET%2F38.3-4.330>
- West, S. G., Wu, W., McNeish, D., & Savord, A. (2023). Model fit in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 184-205). Guilford Press.
- Whittlesey, V., & Steiner, H. H. (2021). The strategy project: An exploration of enhancing self-regulated learning in an introductory psychology course. *InSight: A Journal of Scholarly Teaching*, 16, 33-54. <https://doi.org/10.46504/16202103wh>
- Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, 8(4), 327-353. <https://doi.org/10.1016/S1041-6080%2896%2990022-9>
- Winne, P. H. (1997). Experimenting to bootstrap self-regulated learning. *Journal of Educational Psychology*, 89(3), 397-410. <https://doi.org/10.1037/0022-0663.89.3.397>
- Winne, P. H. (2005). A perspective on state-of-the-art research on self-regulated learning. *Instructional Science*, 33, 559-565. <https://doi.org/10.1007/S11251-005-1280-9>
- Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist*, 41(1), 5-17. https://doi.org/10.1207/s15326985ep4101_3
- Winne, P. H. (2010). Bootstrapping learner's self-regulated learning. *Psychological Test and Assessment Modeling*, 52(4), 472-490.
- Winne, P. H. (2013). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 145-178). Routledge.
- Winne, P. H. (2017a). Cognition and metacognition within self-regulated learning. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 36-48). Routledge. <https://doi.org/10.1093/acrefore%2F9780190264093.013.1528>
- Winne, P. H. (2017b). Learning analytics for self-regulated learning. In C. Lang, G. Siemens, A. Wise, & D. Gasevic (Eds.), *Handbook of learning analytics* (pp. 241-249). Society for Learning Analytics Research. <https://doi.org/10.18608/HLA17.021>

- Winne, P. H. (2021). Cognition, metacognition, and self-regulated learning. In D. S. Dunn (Ed.), *Oxford research encyclopedia of education*. Oxford University Press.
- Winne, P. H. (2022). Modeling self-regulated learning as learners doing learning science: How trace data and learning analytics help develop skills for self-regulated learning. *Metacognition and Learning*, 17(3), 773-791. <https://doi.org/10.1007/s11409-022-09305-y>
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9781410602350-19>
- Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology*, 27(4), 551-572. <https://doi.org/10.1016/S0361-476X%2802%2900006-1>
- Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, 51, 409-428. <https://doi.org/10.3758/s13428-018-1055-2>
- Xiao, Y., Liu, H., & Hau, K.-T. (2019). A comparison of CFA, ESEM, and BSEM in test structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(5), 665-677. <https://doi.org/10.1080/10705511.2018.1562928>
- Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. University of California, Los Angeles.
- Zhang, B., Luo, J., Sun, T., Cao, M., & Drasgow, F. (2023). Small but nontrivial: A comparison of six strategies to handle cross-loadings in bifactor predictive models. *Multivariate Behavioral Research*, 58(1), 115-132. <https://doi.org/10.1080/00273171.2021.1957664>
- Zimmerman, B. J. (2013). From cognitive modeling to self-regulation: A social cognitive career path. *Educational Psychologist*, 48(3), 135-147. <https://doi.org/10.1080/00461520.2013.794676>
- Zimmerman, B. J. (2015). Self-regulated learning: Theories, measures, and outcomes. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 21, pp. 541-546). Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.26060-1>
- Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, 53(1), 141-160.
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70, 123-133. <https://doi.org/10.1007/s11336-003-0974-7>

Addendum: Complementary Person-centered Validation

Purpose

A complementary person-centered analysis was conducted to further investigate the multidimensionality of the construct. This analysis compared latent profiles based on specific factors alone with latent profiles that included both general and specific factors. The goal was to test how the addition of the general factor alters the scale's ability to segment the sample into meaningful clusters of participants. Given the diagnostic objectives of this scale, this approach was deemed beneficial. Additionally, the profiles derived from the symmetrical and S-1 bifactor models were compared to examine the impact of anchoring the general factor on metacognitive challenges within the extracted profiles. Finally, the associations of extracted profiles were explored with three criterion variables: GPA expectations, GPA self-efficacy, and end-of-semester GPA. This exploration aimed to understand how these latent clusters relate to academic outcomes, providing insight into the practical diagnostic objectives of the scale.

Background

A person-centered modeling can be employed as a complementary analytical method to further examine the multidimensional structure of the academic challenges scale. Person-centered approaches offer a unique perspective, focusing on the relationships among individuals rather than variables.

Variable-centered modeling aims to understand how variables relate to each other across a sample of individuals. In other words, the primary focus of this approach is to discern patterns of relationships among variables. It operates under the assumption that these relationships are

relatively consistent across individuals in the sample, and thus, the characteristics of the sample can be summarized by average estimations drawn from the sample (Howard & Hoffman, 2018).

Unlike the variable-centered approach, which assumes that participants are drawn from a relatively homogeneous population and can thus be represented by the average estimates of the sample parameters, the person-centered approach views the population as a mixture of distinct latent clusters, requiring a few cluster-specific estimates to represent the data structure (Howard & Hoffman, 2018). The primary interest in person centered analyses such as latent profile analysis is to identify clusters of individuals who exhibit similar characteristics which differentiate them from other clusters (Collins & Lanza, 2009; Magidson et al., 2020). These clusters are often referred to as "latent" classes or profiles because they are not directly observed but are estimated from the indicators after partialling out the measurement error.

Nevertheless, person-centered modeling is not typically employed as a method for investigating the psychometric adequacy of constructs. This may be due to the perception that it represents a rival epistemological approach to the variable-centered method. The psychometric quality of a construct often involves exploring the reliability and psychometric validity of the constructs (variables) under investigation, rather than understanding how the measure can identify distinct clusters within the population.

However, while Latent Profile Analysis (LPA) and Common Factor Analysis (e.g., CFA) stem from different epistemological perspectives and are used to answer different types of research questions, the underlying computations share many similarities. Regarding modeling the covariance structure, an LPA in which individuals are grouped into K profiles based on their pattern of responses across a set of observations is mathematically equivalent to a common factor analysis in which the covariance among the set of observations is explained by a $K-1$ number of underlying latent factors (Bauer & Curran 2004, Kankara, et al., 2011).

Similarly, LPA and factor analysis are equivalent in the sense that they both partition the covariance among the observations into a set of underlying dimensions (either profiles or factors). However, they do this in different ways and make different assumptions about the nature of these dimensions. Both LPA and CFA model indicators as continuous observations. However, LPA assumes that the latent dimensions are categorical (i.e., distinct profiles), while the CFA assumes that the dimensions (i.e., factors) are continuous (Magidson et al., 2020); That is why despite such a mathematical equivalence in terms of their implications for the covariance structure, they may not be equivalent in terms of their interpretability or suitability for a particular research question or data.

Understanding how the population can be decomposed into distinct subpopulations in terms of the subscales of a construct is beneficial for exploring its multidimensionality, particularly when it includes an overarching general factor (Morin et al., 2017). Toward this aim, by comparing two person-centered models, one with first-order factors as indicators of latent profiles and one including both specific and general factors as indicators of latent profiles, we can gain a better understanding of the extent to which the identified profiles predominantly differ in terms of the configurations of the profile indicators, or whether they represent similar configurations that only differ quantitatively in terms of level (low, medium, high). These two types of profile differentiation are referred to as shape-differentiated and level-differentiated, respectively (Morin et al., 2017). Observing the changes in shape and level characteristics of profiles when the general factor is introduced to the specific factors helps in understanding the multidimensionality of the scale. Since the general factor represents the common variance among items, it is assumed that once the general factor is added, the level differentiation will be captured by the general factor, allowing further shape differentiation by the specific factors.

If the latent profile analysis identifies profiles that are predominantly differentiated by level differences, the sample can be effectively treated as a homogeneous pool, which can be adequately modeled using variable-centered approaches. If meaningful shape differentiation

exists, interpreting an individual's score on a variable (e.g., one's score on motivation challenge) should be contextualized within the profile to which they belong. Similarly, the strong shape differentiation across the factors may suggest a more complex interaction among the factors than what can be captured by a covariation structure represented by a general factor.

Analytic Strategy

Conducting fully latent profile models often struggles to converge on a proper solution, particularly when the number of indicators is large (Collins & Lanza, 2009; Magidson et al., 2020). As a result, the analysis of mixture models is usually based on summed or average items. In this study, factor scores from the best-fitting models were used as indicators of profiles. While factor scores do not completely account for measurement errors like a fully latent model would, however, they are assumed to partially account for measurement error because they assign higher weights to observations with lower measurement error. In this study, factor scores were computed using the Maximum A Posteriori (MAP) method (Skrondal & Laake, 2001), which identifies the maximum of the posterior distribution of the factor. The deviation between factor scores and true factors is a function of the reliability of items within a factor and can be measured by the factor determinacy index (Rodriguez et al., 2016) discussed earlier in the reliability section.

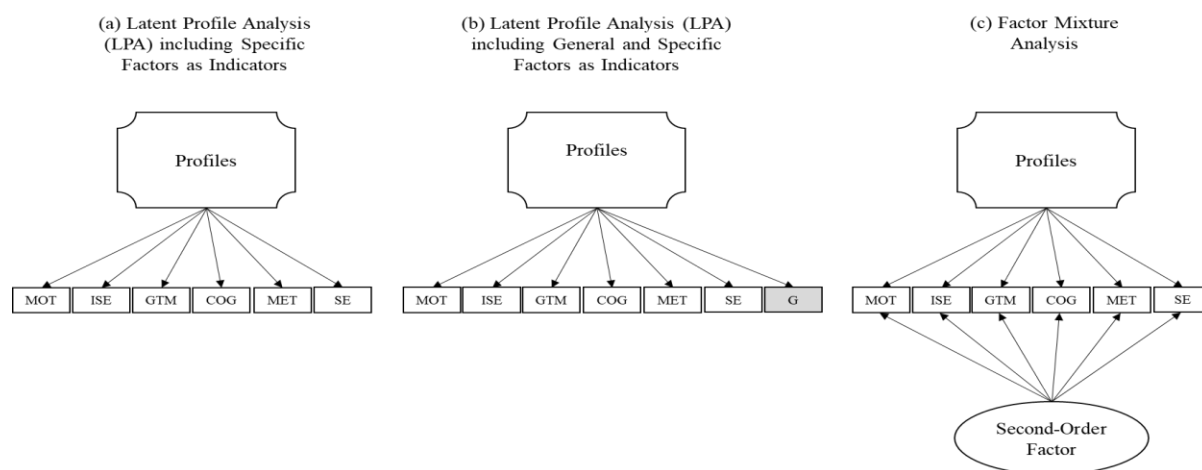
Two latent profile analyses shown in Figure 1a and 1b were used to investigate the multidimensionality of the academic challenges scale. The first LPA used factor scores from the best-fitting first-order model (e.g., First order ESEM) as indicators of profiles. The second used the best-fitting bifactor model (Bifactor γ_{-1} ESEM) factor scores as indicators of profiles. As discussed earlier, in higher-order constructs, the first-order and higher-order constructs are not appropriately disaggregated and share overlapping variance (Chen et al., 2006). Therefore, this step is only feasible using a comparison of a first-order and a bifactor solution.

Comparing these two models (1a and 1b) provides insight into the degree to which the inclusion of the general factor could absorb level differences across profiles, facilitating

additional qualitative differentiation of profiles in the second model (1b) compared to the first (1a). In other words, the inclusion of the general factor should result in profiles in which general factor represent a considerable amount of quantitative differentiation between profiles (level), while specific factors would represent more shape-differentiation (Morin et al., 2017).

Figure 1.

Person-Centered Mixture Models for the Academic Challenges Scale



Note. Abbreviations: MOT: Motivation, ISE: Initiating & Sustaining Engagement; GTM: Goal & Time Management; COG: Cognitive, MET: Metacognitive Challenges; SE: Social & Emotional Challenges.

Moreover, such a comparison will aid in determining to what extent specific factors retain their anticipated meaning, rather than reflecting method effects, following the addition of the global factor. This thus contribute to our understanding of the scales' multidimensionality (Morin et al., 2017).

Another potential person-centered alternative for LPA is Factor Mixture Modeling (Figure, 1c) which was not used in this study for a couple of reasons. Firstly, this model should include the factor scores from the best-fitting first-order factor model, with a latent general factor added to account for the commonalities among profile indicators. However, this general factor does not include all the covariance among indicators, but only the residual covariance left unexplained by the latent profiles. More importantly, for the sake of model identification, the

mean of this general factor must be constrained to equality among profiles. This assumption limits its usefulness for investigating the impact of using coexisting general and specific factors as indicators of profiles. In other words, although it allows for partial disentanglement of general and specific factors, it assumes that all profiles are invariant in terms of the level of the general factor (Morin et al., 2017; Morin & Marsh, 2015).

The Criteria for Deciding which Profile Solution to Retain

The criteria for retaining the profile solutions are provided in Table 1.

Table 1

The criteria used for deciding which latent profiles to retain

<i>Theoretical Adequacy of the Solution</i>	
Profiles should be theoretically meaningful, stable and reliable (Collins & Lanza, 2009; Marsh et al. 2009; Magidson et al., 2020)	
Profiles' sizes should make sense	
<i>Statistical Adequacy of the Solution</i>	
• absence of negative variance estimates; (Bauer & Curran 2004) • model convergence on a replicated solution (indication of real maximum likelihood) • means and variances would be freely estimated across profiles unless an improper solution or convergence issues arise due to overparameterization. In such cases, variances are set to be equal across profiles.	
<i>Index</i>	<i>Criteria</i>
Akaike Information Criterion (AIC)	A lower value on the AIC, CAIC, BIC, and ABIC suggests a better-fitting model
Consistent AIC (CAIC)	
Bayesian Information Criterion (BIC)	
sample-size Adjusted BIC (ABIC)	
LO-MENDELL-RUBIN ADJUSTED LRT TEST (aLMR)	
Bootstrap Likelihood Ratio Test (BLRT)	• The aLMR and BLRT compare a k-class model with a $k-1$-class model. A significant p value indicates that the $k-1$-class model should be rejected in favor of a k-class model. • heavily influenced by sample size (Marsh et al. 2009), In large sample size, interpret in the context of the elbow plot
Elbow Plots representing information criteria	the point after which the slope flattens indicates the optimal number of profiles.
Entropy	Ranging from 0 to 1, with higher values indicating less classification errors. should not be used to determine the optimal number of profiles (Lubke and Muthén 2007).

Note. References: (Bauer & Curran, 2004; Hipp & Bauer, 2006; Nylund et al., 2007; Peugh & Fan, 2013; Tein et al., 2013; Tofighi & Enders, 2008; Tolvanen, 2007)

Analytic Procedure for Computing Mean Differences of Criterion Variables Across Profiles

We explored mean differences in criterion variables across the identified profiles to better understand the scales' capacity to identify students who may benefit most from screening and require additional support. This exploration is particularly relevant given that the data were collected during the initial weeks of the academic semesters.

To investigate the differences between profiles, Analysis of Variance (ANOVA) was used. Considering that ratings of GPA self-efficacy were anchored to students' GPA expectations, Analysis of Covariance (ANCOVA) was employed to control for variations in GPA expectations when examining differences in GPA self-efficacy across profiles. Given that the assumptions underlying these methods might not always be met, alternative approaches such as the Brown-Forsythe (Brown & Forsythe, 1974) and Welch ANOVA (Welch, 1951) were used. These two alternatives are designed to handle situations where the assumption of homogeneity of variances is violated. The Brown-Forsythe test, which assesses differences in the median ranks among groups, is robust to non-normal or skewed data distributions and the presence of outliers. Welch's test adjusts the degrees of freedom based on sample size and variance, offering a more accurate Type I error rate compared to traditional ANOVA when variances between groups are not homogenous (Delacre et al., 2019).

To manage the family-wise error rate, conservative methods like the Bonferroni correction were compared with approaches such as the Holm, Games-Howell, and Dunn tests. While Bonferroni can lead to Type II errors, failing to detect real differences, the Holm-Bonferroni correction is a step-down procedure that provides a balance between maintaining statistical power and controlling for false discoveries (Abdi, 2010). The Games-Howell test, which does not assume homogeneity of variances, is particularly suited for situations where variances and sample sizes are unequal across groups (Games & Howell, 1976). Additionally,

the Dunn test was used for non-parametric pairwise comparisons following a Kruskal-Wallis test, helping to control for Type I error across multiple tests (Dinno, 2015).

Reporting the statistical significance of differences among profiles using these various methods facilitated a comprehensive assessment of Type I and Type II errors and ensured more robust and reliable conclusions.

Results

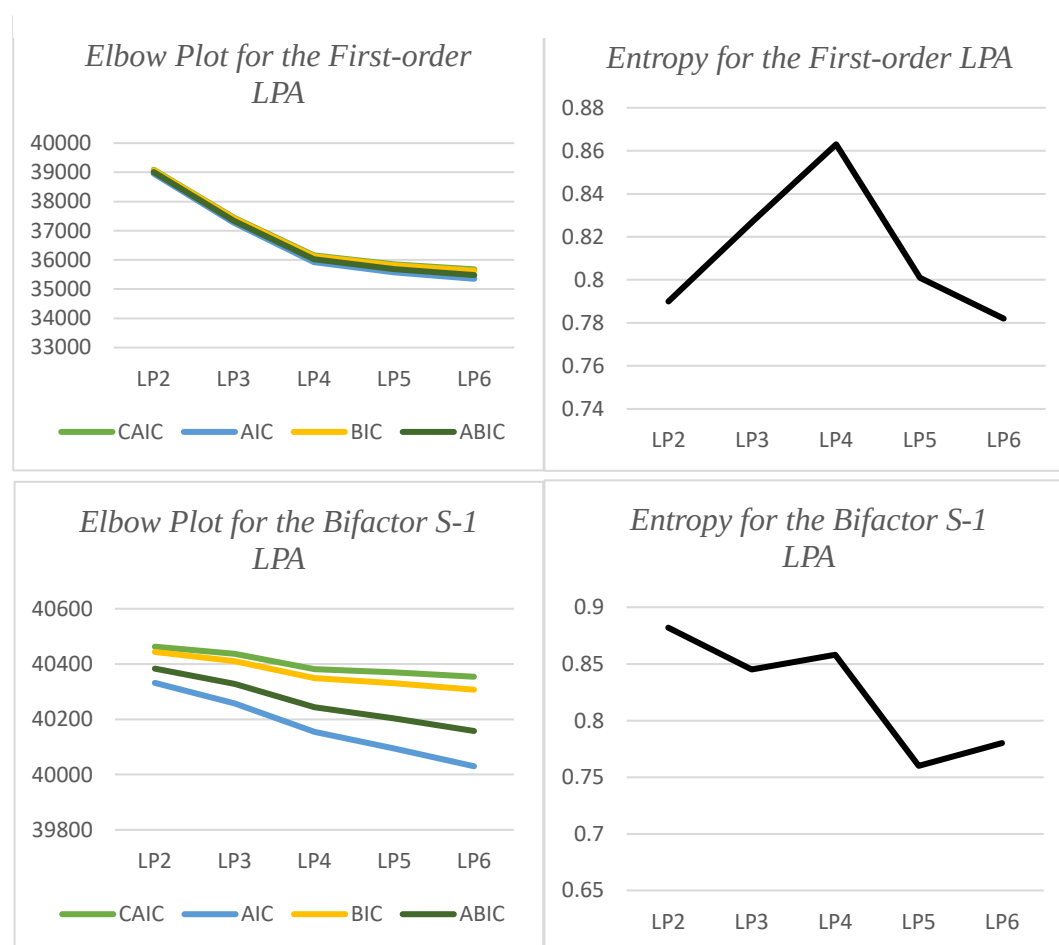
Latent Profile Analyses with equal variances across profiles were conducted to further investigate the multidimensionality of the academic challenges scale using a person-centered approach. Models with free variances across variances are provided in the supplementary materials. The decision to use models with equal variances across groups was made to avoid local maxima, which demanded increased starting values when variances were freely estimated across groups. Therefore, the more parsimonious models with equal variances across groups were selected for their stability. These models also showed better classification accuracy (entropy) compared to the analyses where variances were freely estimated across the profiles. LPA analyses were conducted using MLR estimator, 5,000 random starts, 500 iterations for these random starts, and the 250 best retained for final stage optimization. All reported models converged on a replicated solution and can be assumed to reflect a “real” maximum likelihood.

Factor scores from the first order ESEM and bifactor ESEM models were used as the indicators of profiles. For the LPA model which used the indicators of the 6-factor ESEM model, both adjusted Lo-Mendell-Rubin likelihood ratio test (Table 2) and elbow plot (Figure 2) supported extracting 4 profiles. For the LPA model which used the factor scores from the bifactor S-1 ESEM model as its indicators, LRT indicated that no profiling was needed but elbow plot showed that a 4-profile solution might be appropriate.

Table 2
Fit Indicators for the Latent Profile Analyses

Fit Indicators for the Latent Profile Analyses							
	CAIC	AIC	BIC	ABIC	Entropy	ADJUSTED LO-MENDELL-RUBIN LRT TEST	
6 factor ESEM							
LP2	39088.8	38957.8	39069.8	39009.4	0.79	3743.9	0.0
LP3	37452.3	37273.1	37426.3	37343.7	0.83	1668.5	0.0
LP4	36157.2	35929.8	36124.2	36019.4	0.86	1333.2	0.0
LP5	35851.6	35576.0	35811.6	35684.5	0.80	361.3	0.0
LP6	35677.3	35353.4	35630.3	35480.9	0.78	232.4	0.4
Bifactor ESEM S-1							
LP2	39770.7	39598.5	39745.7	39666.3	0.59	876.7	0.0
LP3	39495.2	39233.3	39457.2	39336.4	0.68	387.4	0.0
LP4	39427.5	39076.0	39376.5	39214.4	0.69	181.5	0.0
LP5	39301.0	38860.0	39237.0	39033.7	0.66	158.0	0.2

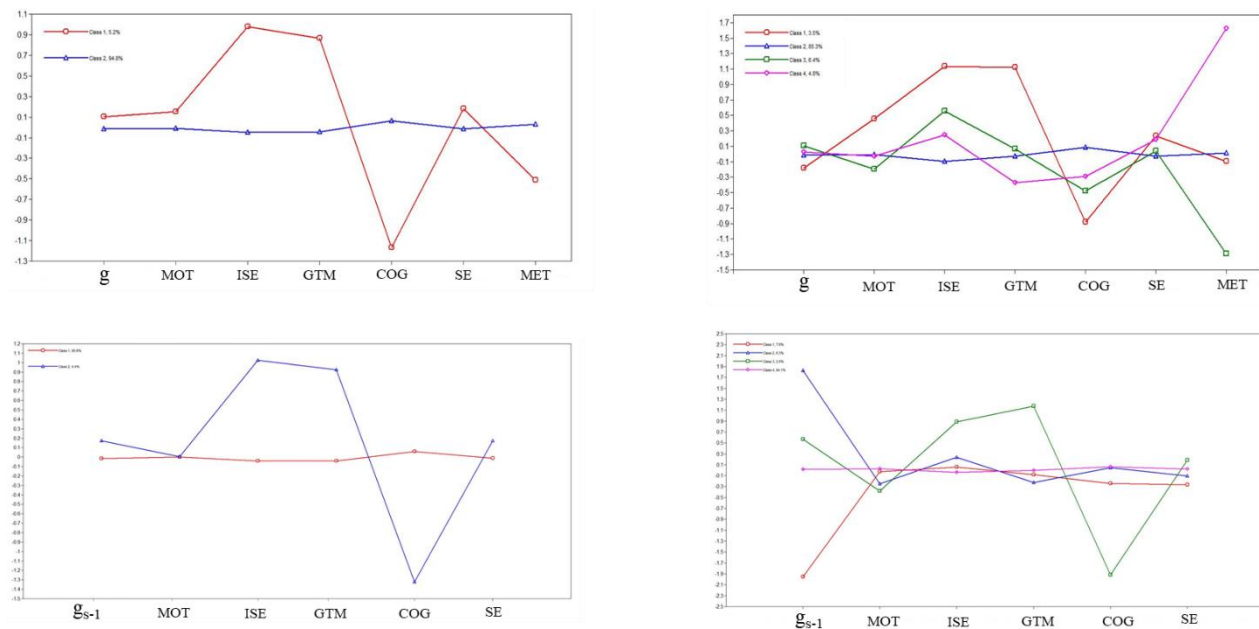
Figure 2



Bifactor ESEM vs. S-1 Bifactor ESEM. An analysis of the best fitting solutions from the S-1 bifactor ESEM model (2 profiles) and the symmetrical bifactor ESEM model (4 profiles), as depicted in Figure 3, revealed similar profile patterns across both models. The primary distinction was observed in the modeling of the metacognitive scale; profiles did not differ in terms of the general factor in the symmetrical bifactor model when the metacognitive scale was included. However, in a S-1 bifactor model, when the metacognitive scale was omitted as an indicator of a specific factor, the g_{s-1} factor absorbed similar differences that the metacognitive scale represented in the symmetrical model. In line with the main analyses, this observation also suggested redundancy between the metacognitive and the general factors and supports the superiority of the bifactor S-1 ESEM model.

Figure 3

Extracted Profiles across Bifactor ESEM and S-1 Bifactor ESEM



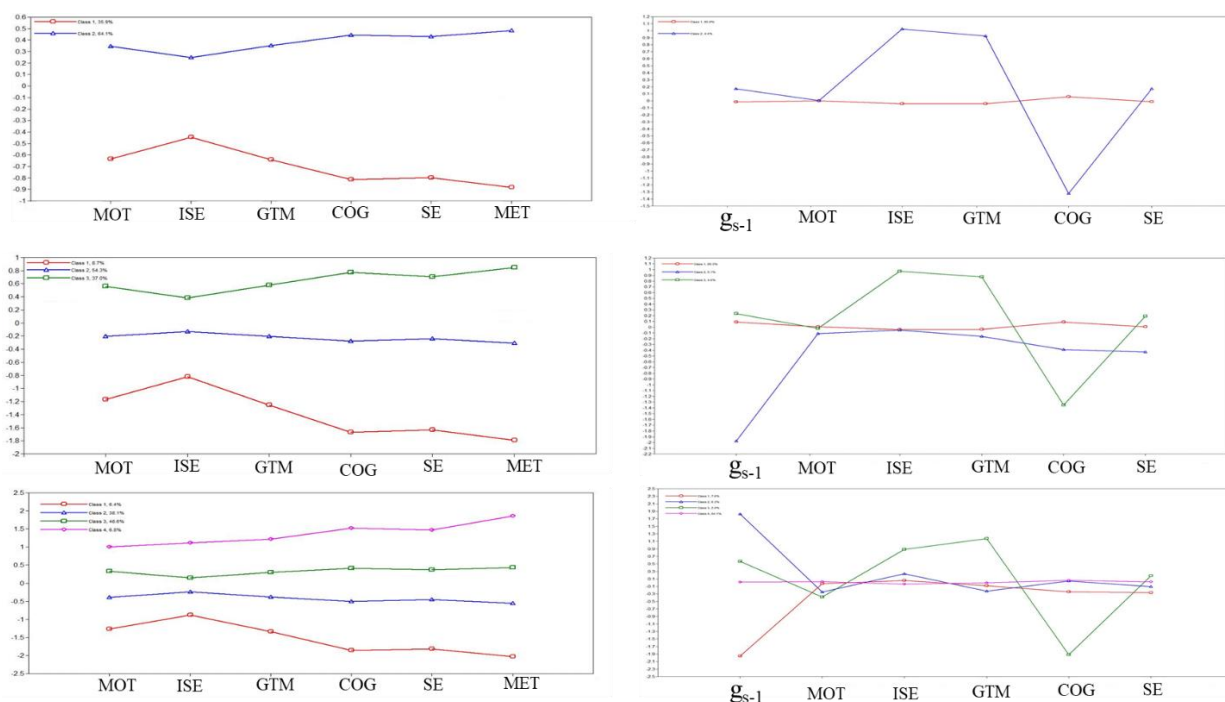
Note. The top row represents the extracted profiles when the factor scores from the symmetrical bifactor ESEM were used as the indicators of the profiles. The bottom row represents the extracted profiles when the factor scores from the S-1 bifactor ESEM were used as the indicators of the profiles. 2 and 4 profile solutions are depicted.

First Order ESEM vs. S-1 Bifactor ESEM. Examining the profiles extracted from the scores of the first-order ESEM (Figure 4, first column) revealed that the distinctions across profiles are entirely based on level differentiation. When the general factor is added, this level differentiation is absorbed by the general factor, and the specific factors then exhibited qualitative shape differences within profiles (Figure 4, second column). It should be noted that interpreting profiles derived from the S-1 bifactor model must be approached with caution due to prior analyses indicating low reliability and low factor determinacy for some specific scales (e.g., GTM, MOT).

In conclusion, the current person-centered analyses underscored the importance of modeling a general factor and supported the approach of anchoring this general factor in the metacognitive items.

Figure 4

The Extracted Profiles across First-order ESEM and Bifactor S-1 ESEM Models



Note. The left column represents the extracted profiles when the factor scores from the first-order ESEM were used as the indicators of the profiles. The right column represents the extracted profiles when the factor scores from the bifactor S-1 ESEM were used as the indicators of the profiles.

Criterion Validity of Profiles

As the final step, the association of profiles with criterion variables was examined to determine if there are statistically significant mean differences across the profiles in terms of expected GPA, GPA self-efficacy, and end-of-semester GPA. The complete parameter estimates for the ANOVA and ANCOVA analyses are available in the supplementary material. For a description of the tests conducted to address potential violations of these tests, please refer to the analytic strategy section of the addendum.

The mean differences across the profiles are illustrated in Figure 5. In these plots, profiles are numerically labeled according to the global level of challenges they represent. For example, Profile 1 is characterized by the lowest level of challenges, whereas Profile 4 is characterized by highest level of challenges. In the first-order ESEM profiles, this level differentiation is represented by quantitative differences in the level of all specific factors. In the bifactor S-1 models, such a level differentiation is based on the level of the g_{s-1} factor.

As the plots show, significant differences across profiles can generally be explained by these level differences in academic challenges, regardless of the shape differences across the profiles.

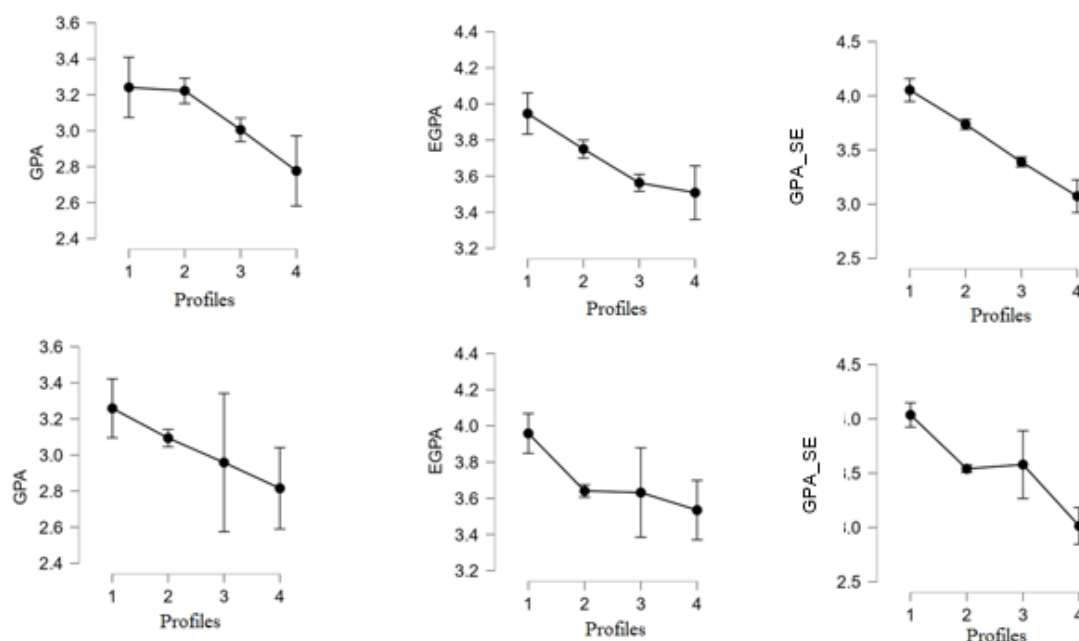
Across the models shown in Figure 4, the majority of respondents (about 85%) were clustered in profiles characterized by middle levels of academic challenges, while approximately 6 to 7 percent of respondents were characterized by high levels of challenges and another 6 to 7 percent were characterized by low levels of challenges. The difference between the profiles characterized by the lowest levels of challenges and the highest was about half a standard deviation. The standard deviation of GPA was .9 indicating a qualitative level shift (e.g., from

GPA A to B or E to F). Statistically significant differences in terms of GPA expectation and GPA self-efficacy were observed across almost all the profiles.

To wrap up, in line with prior analyses, the person-centered analysis supported that the differences across profiles are predominantly level differences which should be represented by a general factor in a variable centered measurement model. Such a predominantly level differentiated indicated the sufficiency of the variable centered analysis for modeling the academic challenges scale.

Figure 5

Mean Differences of Criterion Variables Across Profiles



Note. The top row represents the mean differences across the profiles derived from the first-order ESEM model.

The bottom row represents the mean differences across the profiles derived from the bifactor S-1 ESEM model.

Vertical error bars represent 95% confidence intervals. For example, Profile 1 is characterized by the lowest level of challenges, whereas Profile 4 is characterized by highest level of challenges.

Discussion

This addendum included a complementary person-centered analysis which involved comparing two Latent Profile Analysis (LPA) models to delve deeper into the multidimensionality of the academic challenges scale. The two models compared were: (a) one using scores from the best-fitting first-order model as indicators for the profiles (termed first-order LPA) and (b) one where scores from the best-fitting bifactor model served as indicators for the profiles (termed bifactor LPA).

Results from the first-order LPA revealed that the extracted profiles were characterized mainly by quantitative level differences among the factors. For instance, different profiles exhibited consistently low, medium, or high levels across all challenge areas. Put differently, these profiles showed no significant qualitative configurations or shape differentiation among the factors, indicating a lack of distinct factor-specific patterns that diverged in shape across profiles.

In contrast, the bifactor LPA, which incorporated both general and specific factors as indicators, demonstrated that level differences among factors were effectively absorbed by the general factor allowing the specific factors to represent more nuanced shape differentiations.

In line with the results of the main analysis, LPA models derived from the symmetrical and S-1 bifactor models also suggested psychometric redundancy of the general and metacognitive factors and supported the superiority of the S-1 bifactor model.

However, it is important to note that the findings from the bifactor LPA should be approached with caution due to the low reliability and low factor score determinacy of the specific factors. This limitation suggests that the specific factors may not have been measured with sufficient precision or consistency, potentially affecting the robustness of the profiles derived in this analysis.

Despite this, the primary purpose of conducting this clustering analysis was to explore the structural complexities that can threaten the appropriateness of using traditional variable-centered analyses for modeling the multidimensionality of the construct. Using person-centered analyses as a complementary method can reveal the degree of oversimplification of the complexity that forcing a one-size-fits-all model may have created onto a sample of potentially heterogeneous groups. To elaborate, the existence of complex shape differentiation among the first-order factors, if observed, would pose a challenge to the validity of interpreting the general factor as a representation of the covariation of observations and factors. Such a scenario would suggest the need for more complex or even non-linear structures to adequately model what the dimensions of a construct represent in combination. However, the level differentiation among first-order factors in this study indicated that the relationships between indicators could be effectively modeled using variable-person centered analyses, without necessarily resorting to mixture modeling such as latent profile analysis.

Furthermore, there were significant mean differences in terms of the criterion variables across profiles. In the first-order LPA, these differences were observed between profiles that consistently differed in terms of levels of all six challenges, while in the bifactor S-1 LPA, the significant differences were observed between profiles that primarily differed in terms of levels of the general S-1 factor.

Overall, these findings highlight the effectiveness of bifactor modeling in modeling the complex interplay of multiple dimensions into a more manageable and interpretable structure.

As the scale is intended for diagnostic purposes, future studies should explore the practical significance of the differences among the extracted profiles. Research should determine which levels of challenge or patterns within these profiles are critical and necessitate

intervention. Additionally, these findings should be translated into observable units of measurement for practical application. Understanding at what point and under what conditions support should be provided is essential for the effective use of the scale in educational settings.

This complementary approach should be viewed as an exploratory investigation into the multidimensionality of the construct using a person-centered approach. Future studies should consider using clustering methods to identify specific groups of struggling students who could benefit from tailored support strategies. The practical recommendations for clustering individuals based on their academic challenges scores are provided in Table 3.

Table 3

Practical Recommendations for Clustering Individuals Based on their Academic Challenges Scores

Is Mixture modeling/ SEM available?	
Yes	• Use LPA with specific factors as indicators (no need to include g, it is implied in the level-differentiation across profiles) OR • Use norm-based or comparative cut-offs for clustering based on the total score
NO	• Use norm-based or comparative cut-offs for clustering based on the total score

Reference for the Addendum

- Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of Research Design*, 1(8), 1-8. <https://doi.org/10.4135/9781412961288.n178>
- Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9(1), 3-25. <https://doi.org/10.1037/1082-989X.9.1.3>
- Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16(1), 129-132. <https://doi.org/10.1080/00401706.1974.10489158>
- Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41(2), 189-225. https://doi.org/10.1207/s15327906mbr4102_5
- Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.
- Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of Welch's F-test instead of the classical F-test in one-way ANOVA. *International Review of Social Psychology*, 32(1), 13. <https://doi.org/10.5334/IRSP.198>
- Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal*, 15(1), 292-300. <https://doi.org/10.1177/1536867X1501500117>
- Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1(2), 113-125. <https://doi.org/10.3102/10769986001002113>
- Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, 11(1), 36-53. <https://doi.org/10.1037/1082-989X.11.1.36>
- Howard, M. C., & Hoffman, M. E. (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods*, 21(4), 846-876. <https://doi.org/10.1177/1094428117744021>

- Kankara, M., Vermunt, J. K., & Moors, G. B. D. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, 40(2), 279-310.
<https://doi.org/10.1177/0049124111405301>
- Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(1), 26-47. https://doi.org/10.1207/S15328007SEM1401_2
- Magidson, J., Vermunt, J. K., & Madura, J. P. (2020). Latent class analysis. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), *SAGE Research Methods Foundations*. SAGE Publications Limited. <http://dx.doi.org/10.4135/9781526421036>
- Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(2), 191-225. <https://doi.org/10.1080/10705510902751010>
- Morin, A. J., Boudrias, J. S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., Madore, I., & Litalien, D. (2017). Complementary variable-and person-centered approaches to the dimensionality of psychometric constructs: Application to psychological wellbeing at work. *Journal of Business and Psychology*, 32(4), 395-419.
<https://doi.org/10.1007/s10869-016-9448-7>
- Morin, A. J., & Marsh, H. W. (2015). Disentangling shape from level effects in person-centered analyses: An illustration based on university teachers' multidimensional profiles of effectiveness. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 39-59.
<https://doi.org/10.1080/10705511.2014.919825>
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 535-569.
<https://doi.org/10.1080/10705510701575396>
- Peugh, J., & Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: A Monte Carlo simulation. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(4), 616-639. <https://doi.org/10.1080/10705511.2013.824780>
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137-150.
<https://doi.org/10.1037/met0000045>

- Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, 66(4), 563-575. <https://doi.org/10.1007/BF02296196>
- Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(4), 640-657. <https://doi.org/10.1080/10705511.2013.824781>
- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in Latent Variable Mixture Models* (pp. 317-341). Information Age Publishing.
- Tolvanen, A. (2007). *Latent growth mixture modeling: A simulation study* (Doctoral dissertation No. 111). University of Jyväskylä.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4), 330-336. <https://doi.org/10.1093/BIOMET%2F38.3-4.330>

Supplementary Materials for Study 3

The Academic Challenges Items can be found on page 465 of the dissertation, while the Descriptive Statistics are available on page 466. The online supplementary materials for Study 3 are accessible at <https://osf.io/bka42/>.