Performance-based Measures of Executive Function and BRIEF-P in Preschoolers: A Latent Variable Approach

by

Yaewon Kim

B.Sc., University of Victoria, 2017

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Psychology

I acknowledge with respect the Lekwungen peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

Performance-based Measures of Executive Function and BRIEF-P in Preschoolers:

A Latent Variable Approach

by

Yaewon Kim

B.Sc., University of Victoria, 2017

Supervisory Committee
Dr. Ulrich Müller, Supervisor
Department of Psychology

Dr. Michael Masson, Departmental Member
Department of Psychology

Dr. John Sakaluk, Non-unit Member
Department of Psychology, Western University

Abstract

Preschool years are an important period for executive function (EF) development. The two common ways of assessing EF in preschoolers are performance-based (PB) and rating measures. One of the most commonly used rating scales for preschoolers is the Behaviour Rating Inventory of Executive Function – Preschool Version (BRIEF-P). The current study explored the longitudinal relationship between three PB measures (Grass/Snow, Shape School, Self-ordered Pointing) and corresponding BRIEF-P scales (Inhibit, Shift, Working Memory) in typically developing preschoolers. There were three assessments in six-month intervals. Participants included 101 children at Time 1, with 86 and 75 in subsequent assessment time points. Using a latent variable approach, longitudinal measurement invariance was tested, supporting partial strong invariance. Results showed a lack of direct correlations between PB measures and corresponding BRIEF-P scales across time. These findings were interpreted in the context of existing literature, yielding a more nuanced understanding of what these two types of measures assess. Specifically, it is proposed that BRIEF-P measures children's *subjective, average* level of EF, while PB tasks measure their *objective, in-the-moment* EF.

Keywords: executive function, preschool, BRIEF-P, performance-based, assessment

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to extend my deepest thanks to my supervisor, Dr. Ulrich Müller, for his continued guidance and support, encouragement, and patience that allowed me to achieve such a significant academic milestone. I am truly indebted to your mentorship since 2016 when you were my honours thesis supervisor. I would also like to thank Dr. John Sakaluk whose statistical expertise was invaluable in completing this thesis. Thank you, Dr. Michael Masson, for your valuable feedback on my original project and flexibility for the changes that I had to make when the initial plans were disrupted by the unprecedented global pandemic. I am extremely grateful to have had such insightful and understanding committee members. Last but not least, I would like to thank Dr. John Walsh, for his time and consideration in serving as my external examiner.

I would also like to send a special thank you to Dr. Sarah Macoun, and my fellow lab mates, Buse Bedir and Jessica Lewis for your emotional support throughout the program. Thanks should also go to all members of the Child Development Lab for their help with the data collection, as well as all thefamilies who have participated in this study. Finally, special thanks to Dr. Mauricio Garcia- Barrera and Ms. Karen Kienapple, for their gracious and continuous support for my journey through the program.

Performance-based Measures of Executive Function and BRIEF-P in Preschoolers:A Latent Variable Approach

The construct of executive function (EF) refers to higher-order processes involved in the conscious control of cognition, emotions, and behaviours (Friedman & Miyake, 2017; Müller & Kerns, 2015). Several previous studies have examined the structure of EF (e.g., Best & Miller, 2010; Garon et al., 2008; see also Karr et al., 2018, for a review). The current study adopts a widely popular and empirically well-supported account (Miyake et al., 2000), conceptualizing EF as having three fundamental components: inhibition, working memory, and cognitive flexibility. There has been a growing body of research examining EF in preschoolers, driven by findings that the preschool years are an important developmental period during which EF undergoes significant changes (Best & Miller, 2010; Carlson, 2005; Garon et al., 2008) and that individual differences in EF during this period predict many areas of children's subsequent development, including school readiness and academic achievement (Blair et al., 2011; Clark et al., 2010; Fuhs et al., 2014; Willoughby et al., 2017), as well as psychosocial adjustment (e.g., externalizing behaviours – Schoemaker et al., 2013; internalizing behaviours – Nelson et al., 2018; social skill development – Caporaso et al., 2019). Furthermore, difficulties in EF have often been associated with developmental disorders such as Autism Spectrum Disorder (e.g., Craig et al., 2016; Pellicano et al., 2017) and Attention-Deficit/Hyperactivity Disorder (e.g., Willcutt et al., 2005). Taken together, these suggest that it is important to assess preschoolers' EF using valid and reliable measures (Diamond, 2016; Nilsen et al., 2017).

There are two common ways of assessing EF in preschoolers: performance-based (PB) and rating measures (Anderson & Reidy, 2012; Carlson, 2005; Malloy & Grace, 2005 Silver, 2014; Young et al., 2017). Although both PB and rating measures purport to capture EF, previous

studies have reported nonsignificant, if not, weak to moderate correlations between these two types of measures (e.g., Duckworth & Kerns, 2011; Gerst et al., 2017; Miranda et al., 2015; see also Silver, 2014, for more examples). However, most of these studies have been cross-sectional. The purpose of the current study was to extend previous findings by examining the longitudinal relationship between PB and rating measures, the findings of which will help further our understanding of the similarities and differences between these two types of measures in assessing EF among typically developing preschoolers.

Performance-based Measures

PB measures often involve standardized procedures in a structured setting and the assessments usually involve accuracy and/or response time. One of the main advantages of PB measures is that these are designed to assess specific components of EF (e.g., Day/Night task for inhibition; Gerstadt et al., 1993; Dimensional Change Card Sort [DCCS] for flexibility; Zelazo, 2006; see also Anderson & Reidy, 2012; Carlson, 2005, for a review). However, such a conceptual precision is often at the price of conceptual breadth, not fully encompassing the multidimensional and complex nature of EF demanded in everyday life, where multiple EF components are simultaneously required (McCoy, 2019; Silver, 2014). As such, PB measures are often criticized for limited ecological validity. Ecological validity refers to the extent to which the performance on a given task can predict a child's everyday functioning beyond the assessment setting (Franzen & Wilhelm, 1996). In a similar vein, most PB measures are administered in such settings as experimental labs or clinics, where instructions are often standardized, initiated, and scaffolded by examiners who provide children with continuous support, encouragement, and structure with minimal distractions, and thus failing to reflect the very essence of EF used in real-world situations, namely, self- and goal-directedness (Anderson

& Reidy, 2012; Gioia et al., 2002). Furthermore, notwithstanding the often-assumed conceptual precision, task impurity is another concern inherent in these measures (Miyake et al., 2000). Task impurity refers to the situation where a measure taps into processes in addition to what it was designed to assess (Miyake & Friedman, 2012). For instance, children's performance on an inhibition task (e.g., go/no-go task) may be moderated by their working memory, the ability to temporarily store and simultaneously update the implied or given rules in the task (Garon et al., 2008). PB measures may also draw on other non-executive processes such as children's general verbal ability, intellectual/cognitive function, processing speed, and motor function (Barkley, 2012; Chaytor et al., 2006). The task impurity problem may be particularly pertinent to young children including preschoolers, whose EF and non-EF processes are still differentiating and developing (Espy et al., 2016).

Rating Measures

Initially developed to address the issue of ecological validity associated with PB measures, rating measures for preschoolers are often parent- or teacher-rated, where each informant rates children's EF based on their perception or impression of the child's self-regulatory behaviours demonstrated across a range of everyday situations and settings, thus allowing more generalizable perspectives on children's EF (i.e., more ecologically valid; Campbell et al., 2016; McCoy, 2019; Miranda et al., 2015). One of the most commonly used rating scales for preschoolers is the Behaviour Rating Inventory of Executive Function-Preschool Version (BRIEF-P; Gioia et al., 2003). There are other rating measures available for preschoolers, but these are often not specific to EF, assessing rather general behaviours (e.g., Child Behaviour Questionnaire – Putnam & Rothbart, 2006; Child Behaviour Rating Scale - Bronson et al., 1990). Studies suggest that rating measures are also generally less influenced by

other non-executive processes (e.g., no association between BRIEF-P and children's IQ; Mahone & Hoffman, 2007). Nevertheless, inherent to rating measures are the subjectivity and potential biases of informants who may have different sociocultural rules and expectations, which, in turn, may depend on various individual (either informant- or child-specific) and contextual factors (Denckla, 2002; Roth et al., 2014). For example, studies have shown that parent ratings may be influenced by situational factors such as parenting stress, parental frustrations, and negative parenting behaviours (Chen et al., 2017; Gross et al., 2015; Moens et al., 2018). As such, PB and rating measures may differ in the extent to which they are susceptible to various contextual and individual factors. For instance, a recent study showed that children's maltreatment status (e.g., physical abuse and neglect) may moderate the association between PB and BRIEF-P parent reports in that there was less convergence between these two measures among children with more severe maltreatment exposure (Fay-Stammbach & Hawes, 2019).

Children's self-regulatory behaviours may also vary across settings, some of which may be more salient for some informants but not others (Achenbach et al., 2017; Lieberman et al., 2007). Furthermore, the informants' expectations may or may not be developmentally appropriate depending on the extent and nature of previous experiences with the child (Allan et al., 2014). For example, primary caregivers such as parents may have limited experiences with children other than their own, potentially resulting in developmentally inappropriate expectations for their children (Camerota et al., 2018; Isquith et al., 2005; Mccoy, 2019). In contrast, secondary caregivers such as early childhood educators are likely to have experiences with a more diverse and larger group of children, which may inform their judgments of the children's EF (Wolraich et al., 2004). As such, previous studies have reported divergence between parent- and teacher-rated EF (e.g., Schneider et al., 2020), which may also, in part, be attributed to their

relative ecological validity. For example, as compared to parent ratings, teacher ratings may have limited generalizability beyond rather structured settings such as daycares, and thus may be more strongly associated with PB measures (Acar et al., 2019; Miranda et al., 2015; Tamm & Peugh, 2019).

Performance-based Measures vs. Rating Measures

Both PB and rating measures are designed to capture EF; however, previous studies have reported that these two types of measures are often unassociated with each other. For example, a previous meta-analysis showed that of all the correlations between PB and rating measures reported, only 24% were statistically significant, the overall median correlation of which was r = .19 (Toplak et al., 2013). This lack of correlation has been initially explained by the limited ecological validity of PB measures and differences between the contextual demands posed by these two types of measures on children (Anderson & Reidy, 2012; Silver, 2014; Ten Eycke & Dewey, 2016; Toplak et al., 2013).

An alternative account is that these two types of measures tap into different aspects of EF (Toplak et al., 2013; see also, Isquith et al., 2013; McAuley et al., 2010; Ten Eycke & Dewey, 2016). Conceptually, this maps onto how Stanovich (2009, 2011) differentiated the human cognitive ability into the algorithmic and the reflective mind, according to which the first is associated with the efficiency of information processing mechanisms (e.g., working memory and cognitive flexibility), while the latter is concerned with the integration of one's goals and beliefs for successful decision-making processes. In practice, this distinction is closely related to that between optimal (or maximal) and typical performance (Toplak et al., 2013). Optimal performance refers to when children are provided with structured instructions, reflecting how efficiently they process the given information in a task. In contrast, typical performance refers to

when children are not explicitly provided with instructions, and the interpretation of a given task is left open to the children. Thus, it may be that the PB measures draw on the algorithmic mind and assess children's optimal/maximal performance, while rating measures draw on the reflective mind and assess typical performance (Isquith et al., 2013; Mattson et al., 2020; Toplak et al., 2013). This interpretation is further supported by neuroimaging findings. For example, Faridi et al. (2015) reported different neural correlates for PB and rating measures in that children's PB working memory was associated with hippocampal and amygdala volumes while their parent-rated working memory was associated with the cortical thickness of the posterior parahippocampal gyrus, an area associated with contextual learning and memory (see also Mahone et al., 2009).

Furthermore, previous studies suggest that PB and rating measures may have differential predictive utility with regard to children's subsequent developmental outcomes. For instance, Schmitt et al. (2014) previously indicated that while preschoolers' teacher-rated EF and their scores on a PB task were associated with their early mathematics and literacy skills, the former was the strongest predictor of children's literacy skills, with the latter being predictive of mathematics skills. Similarly, Miranda et al. (2015) reported that among 5-6-year-old children, both parent- and teacher-rated BRIEF scales of inhibition and working memory were a stronger predictor of children's ADHD symptomatology, while corresponding PB measures predicted their reading achievement to a greater degree (see also Dekker et al., 2017).

It should be noted, however, that the majority of previous studies have focused on school-aged children and adolescents, using either clinical (Bünger et al., 2019; Davidson et al., 2016; Gardiner et al., 2017; Gross et al., 2015; Kriegar et al., 2018; McAuley et al., 2010; Rai et al., 2017; Tan et al., 2018; Toplak et al., 2009; Vries et al., 2018) or non-clinical samples (e.g.,

Dekker et al., 2017; Gerst et al., 2017; Mahone et al., 2009; Ten Eycke & Dewey, 2016). Relatively few studies have examined preschoolers and most of the available studies in that age range tend to focus on clinical populations (e.g., Daunhauer et al., 2017; Ezpeleta & Granero, 2015; Fay-Stammbach & Hawes, 2019; Jacobson et al., 2018; Loe et al., 2015; Mahone & Hoffman, 2007; O'Meagher et al., 2019). The few studies on typically developing preschoolers have yielded mixed results (e.g., significant associations in Garon et al., 2016; cf. Liebermann et al., 2007). Furthermore, previous studies have often focused on examining the discriminative utility of PB and rating measures in distinguishing clinical groups (e.g., children with ADHD - McAuley et al., 2010; children with ASD – Gardiner et al., 2017; Loe et al., 2015 – pre/full term; Fay-Stammbach & Hawes, 2019 – maltreatment) from typically developing children, reporting correlations for either only the clinical group or the clinical and non-clinical groups combined, with limited studies reporting associations specific to typically developing children. In addition, to my knowledge, no studies to date have examined the extent to which these two types of measures are longitudinally associated. The current study attempted to address this knowledge gap. Specifically, the current study examined: (a) both cross-sectional and longitudinal associations between PB and rating measures; (b) whether these two types of measures assess the same construct; (c) whether children's EF improves over time, and; (d) whether these measurement types differ in the extent to which they are sensitive to children's EF development.

Current Study

The current study had three main research questions. First, I examined both concurrent and longitudinal relationships between PB measures (Grass/Snow, Self-ordered Pointing, Shape School) and corresponding BRIEF-P scales in typically developing preschoolers at both measurement and structural levels. At the measurement level, the current study used correlational

analyses, and consistent with previous literature (e.g., Fuhs & Day, 2011; Garon et al., 2016; Liebermann et al., 2007; Miranda et al., 2015), I hypothesized that there will be nonsignificant direct cross-sectional correlations between these two types of measures. I also examined longitudinal correlations but without specific predictions. To explore whether these two types of measures are concurrently and/or longitudinally associated at the structural level, the current study used latent variable analytic approach. One of the main advantages of such an approach is that it controls measurement errors by extracting only the common variance shared by the PB measures and BRIEF-P scales specified to represent the corresponding latent variables (Little, 2013).

The second research question addressed whether PB measures and BRIEF-P scales assess the same construct by examining whether these two types of measures conform better to a one-factor model (i.e., all PB measures and BRIEF-P scales loaded onto a single latent factor across time). Based on the abovementioned accounts (e.g., algorithmic vs. reflective; maximal vs. typical; Stanovich, 2009, 2011; Toplak et al., 2013), I hypothesized that these two measures are best represented by a two-factor model (i.e., PB measures and BRIEF-P scales loaded onto separate latent factors). Given that the task impurity problem is inherent in both types of measures, I further examined whether PB measures and BRIEF-P scales differ in the extent to which they capture their corresponding latent variables.

The preschool period is a developmental period characterized by a spurt of EF development (Best & Miller, 2010), and thus the availability of developmentally sensitive measures is critical. Previous literature suggests that different EF components show distinct developmental trajectories (see Garon et al., 2008; Müller & Kerns, 2015, for an overview). Yet, there is limited knowledge on how PB measures and BRIEF-P scales differ in the extent to which

they capture the development of children's EF and its components. Therefore, the third research question addressed whether PB measures and BRIEF-P scales assess the same respective latent variables across time by testing longitudinal measurement invariance. By examining whether the same latent variables are measured across time (i.e., factorial invariance), the longitudinal measurement invariance testing can demonstrate whether the observed changes in children's EF over time (if any) are attributable to the actual EF development (Little, 2013; Meredith & Teresi, 2016). It was speculated that there will be different patterns of relations between these measures and corresponding latent variables across the measurement type. Once the longitudinal measurement invariance was established, I examined whether children's EF improved over time and whether one measurement type is better than the other at capturing developmental changes.

Methods

Participants

The sample used for this study was from a longitudinal study exploring language development and EF in preschoolers, with three assessments, each being six months apart on average. The study was conducted in the Child Development Lab at the University of Victoria. Participants were recruited through flyers at local shopping venues, daycares, recreational centres, and community centres in Victoria, BC, and its close surrounding areas (e.g., Esquimalt). To be eligible for the study, children had to be between 36-48 months at the first assessment, have English as their first language with no diagnoses of language impairment or developmental disorders. At the first assessment, a total of 105 preschoolers were recruited, with an aim of collecting 125 participants. Two participants were excluded due to limited verbal ability with suspected language impairment and loss of demographic and parent forms, respectively. Two participants voluntarily withdrew from the study after the first assessment and thus were

excluded. Given the longitudinal design of the study, not all participants were able to complete all three assessments by the time of the current data analysis, resulting in the final sample including 101 participants at the first assessment, 86 at Time 2, and 75 at Time 3. All procedures were approved by the university's institutional ethics board.

Procedure

When recruiting from daycares, the supervisor was asked to provide written consent for recruiting preschoolers. All of the parents of preschoolers were asked to provide written consent for participation upon arrival at the lab. Parents then completed a demographics questionnaire and BRIEF-P (Gioia et al., 2003). At the beginning of testing, each participant was asked to provide verbal assent of his/her participation in the study. Each testing session took about 2.5 hours, consisting of different EF and language measures in a mixed order.

At the end of the assessment, parents received a small monetary compensation ($15, $20, and $25 for each timepoint, respectively). Children were provided with snacks during a short break and stickers as motivational rewards throughout the assessment.

Measures

Not all measures used in the original study are included here. The original study included the measures of theory of mind, language skills, and EF skills that were not of interest for the present study.

Behavioral Rating Inventory of Executive Function – Preschool Version (BRIEF-P)

The BRIEF-P (Gioia et al., 2003) is a parent-, caregiver-, or teacher-report measure of a preschooler's everyday EF behavior, intended for a broad age range of children from 2 years through 5 years 11 months. The measure consisted of 5 scales (emotional control, inhibition, flexibility, working memory, and planning), three of which were included in the present study

(inhibition, flexibility, working memory). Each of the three scales had good internal consistency (Cronbach's α = .77-.90) and acceptable test-retest reliability (ICC = .68-.78). All items were scored on a scale of 1 (never), 2 (sometimes), and 3 (always). BRIEF-P was reverse coded such that a higher score indicates greater difficulties with EF (i.e., lower EF level).

Performance-based Measures

Three performance-based measures were included in thisstudy, assessing inhibition, working memory, and flexibility, respectively.

Grass/Snow (G/S; Carlson & Moses, 2001). The G/S task was used as a measure of inhibition. The materials for this task included two red, felt handprints, two laminated 15 x 15 cm squares: green and white. The handprints were placed directly in front of the child on the table, with the green and white squares placed side by side above the handprints. After confirming children's knowledge of the colors of grass and snow, children were told that this was a silly opposite game, where they should point to the green square when the experimenter said snow and point to the white square when told grass. Children were instructed to return their hands to the top of the red felt handprints after each trial. There were 16 trials in total, each scored as 2 (correct), 1 (self-correction after initially moving towards the incorrect response), or 0 (wrong). The highest possible score was 32. Scores were excluded from data analysis if children did not understand the instructions after three attempts and/or did not know the colors of grass and snow.

Shape School (SS; Espy, 1997). The SS task was used to assess children's flexibility. It had two parts, each of which involved two sheets of paper (21.6 cm x 27.9 cm) displaying shape stimuli. Both parts had a practice (six shapes displayed in two rows) and a test phase (15 shapes displayed in three rows). In the first part, the stimuli used had two dimensions: shape (circle and

square) and color (red, yellow, blue). Children were instructed to name the colors of the shapes the researcher pointed to as fast as possible. In the second part, a new dimension was added such that seven shapes in the test trial (two in practice trial) had glasses drawn on them. For shapes with glasses drawn on them, children were told to name the shapes (instead of colors) as quickly as possible. For the shapes that did not have glasses drawn on them, children were told to name the color as they had done in the first part. The total number of correct responses in the second part was used as the outcome score for this task. The range of possible scores was 0-15. Scores were excluded if children did not pass practice trials after three attempts and/or did not know the shapes (i.e., circle and square) and/or colours (i.e., red, yellow, and blue) included in the task.

Self-Ordered Pointing (SOP; Petrides & Milner, 1982). The SOP task was used to measure children's working memory. Participants were presented with a stimulus booklet containing several sheets of paper (21.6 cm x 27.9 cm). Each page had a set of black and white illustrations on it. The task was introduced as a pointing game, where children were told that they had to point to a picture quietly (i.e., without saying a word) on the first page and to point to a different picture they had not pointed to before on the subsequent pages. The test phase had six levels, each with two trials, and each included a different set of stimuli. Each level differed in the number of items and contained as many pages as there were pictures. The first level included three pictures, and the subsequent levels had sets of 4, 5, 6, 7, and 8 pictures. The layout of the stimuli was consistent throughout each level, while the location of each stimulus changed randomly from page to page. The task was discontinued when children did not successfully complete at least one trial at a level. The outcome variable was the number of successfully completed sets (range = 1-16). Scores were excluded if children verbalized their responses while simultaneously pointing to the pictures even after being explicitly told not to, to account for any

confounding variables (e.g., language ability).

Reliability

Factor loadings presented in the results section were used to compute composite reliability (ω; McDonald, 1999). Composite reliability measures the extent to which the observed indicators are related to the latent variable (Raykov, 1997). The current study used ω as a reliability coefficient because compared to the traditional Cronbach's α, ω does not assume tau equivalence (i.e., equalfactor loadings in manifest variables on the corresponding latent construct). Reliability coefficients of PB measures were weak and variable (ω = .54, .37, .57 for Times 1 through 3, respectively), and lower than those of BRIEF-P scales, which were acceptable, but still variable (ω = .77, .83, .78 for Times 1 through 3, respectively).

Statistical Analysis

All statistical analyses were done using R (R Core Team, 2020). Data screening indicated a few univariate and bivariate outliers; however, with the absence of multivariate[1] outliers, I did not exclude any other data than the ones mentioned above. Both univariate and multivariate normality assumptions were violated[2], and thus robust maximum likelihood (MLR) estimator was used to estimate parameters in all subsequent latent variable analyses. Full information maximum likelihood (FIML) was used to handle missingness in the data.

The main analyses included latent variable analysis including longitudinal invariance testing and sets of confirmatory factor analytic (CFA) model comparisons. Before fitting the initial longitudinal CFA model, I specified an alternative null model with no common factors, with means and variances of the same indicator constrained to equivalence over time (Little,

1. Mahalanobis distances were used to check multivariate outliers with $p < .001$. No multivariate outliers were identified.
2. Multivariate normality assumption was achieved only at Time 2. Multivariate normality assumptions were tested using the "mvn" package, which is not compatible with missing data. Thus, in the current study, missing data were removed when testing multivariate normality assumptions. This resulted in the sample size of 60 at Time 1, 69 at Time 2, and 64 at Time 3, rendering cautious interpretation. All variables showed non-normal univariate distribution.

2013).When fitting the initial longitudinal CFA model, I used the fixed factor scaling method, with latent means fixed to zero and variances to one.

To examine whether PB measures and BRIEF-P scales tap into the same EF latent variable, all manifest variables were loaded onto one single EF latent factor at each time, and this model was compared to the initial longitudinal CFA model (i.e., separate PB and BRIEF factors at each time). Then, to examine whether these two types of measures differ in the extent to which they are explained by their corresponding latent variables, I estimated two additional models with residual variances of manifest variables constrained to equivalence over time within and across factors.

Longitudinal invariance testing was performed to examine whether PB measures and BRIEF-P scales assess the same respective latent variables across time. To test whether individual measures load on to corresponding factors at the same level, a weak invariance was model was estimated (i.e., equivalent factor loadings across time, with latent variances at Times 2 and 3 freely estimated). Next, to test whether the same PB and BRIEF factors are measured across time, a strong invariance model was fit (i.e., equivalent factor loadings and indicator intercepts across time, with latent means and variances at Times 2 and 3 freely estimated). Once the strong factor invariance was achieved, latent mean models were estimated to examine whether children's average EF changed over time. To test which (or whether both) of the PB and BRIEF-P latent means changed over time, two additional models were fit, with respective latent means fixed to equivalence across time.

For all analyses, I used several fit indices to evaluate each model fit. Along with the robust χ^2 difference test, the following additional absolute and relative fit indices were used: Bentler's comparative fit index (CFI), the Tucker-Lewis Index (TLI), and the root-mean-square

error of approximation (RMSEA). The criteria for good model fit based on these indices are CFI/TLI > .95 and RMSEA < .05. However, based on the previous discussion on the reliability paradox (McNeish & Hancok, 2018), factor loadings were taken into account when interpreting these fit indices. Hu and Bentler's (1999) cut-offs are often too liberal for models with weaker factor reliability or too strict for models with stronger factor reliability (as in the current measurement model specified in results sections). Subsequently, for more appropriate model fit comparisons with the alternative null model specified, change in these fit indices were used as criteria for significant loss of fit: ΔCFI/ ΔTLI (decrease) >.01, ΔRMSEA (increase) >.015-.02, and robust $\Delta\chi^2$ with $p < .05$ (see Chen, 2007; Cheung & Rensvold, 2002 for additional discussion).

Results

Using latent variable analysis, I examined the longitudinal association between PB measures and corresponding BRIEF-P scales among typically preschoolers. Table 1 shows participants' demographic information including age, gender, household income level, and proportion of parents with postsecondary education. Overall, children's EF, either assessed by PB measures or BRIEF-P, improved across time albeit at a lesser degree when measured by BRIEF-P (see Figure 1). Intraclass correlations (ICCs) were computed to indicate test-retest reliability conditional on participant age. ICCs were within acceptable levels of .68-.78 for BRIEF-P scales while they were low for PB measure (.28-.33).

Cross-sectional and Longitudinal Correlations

Table 2 shows descriptive statistics and correlations of all the variables included in the study. As expected, individual PB measures were not significantly correlated with corresponding BRIEF-P scales at each time.

Table 1

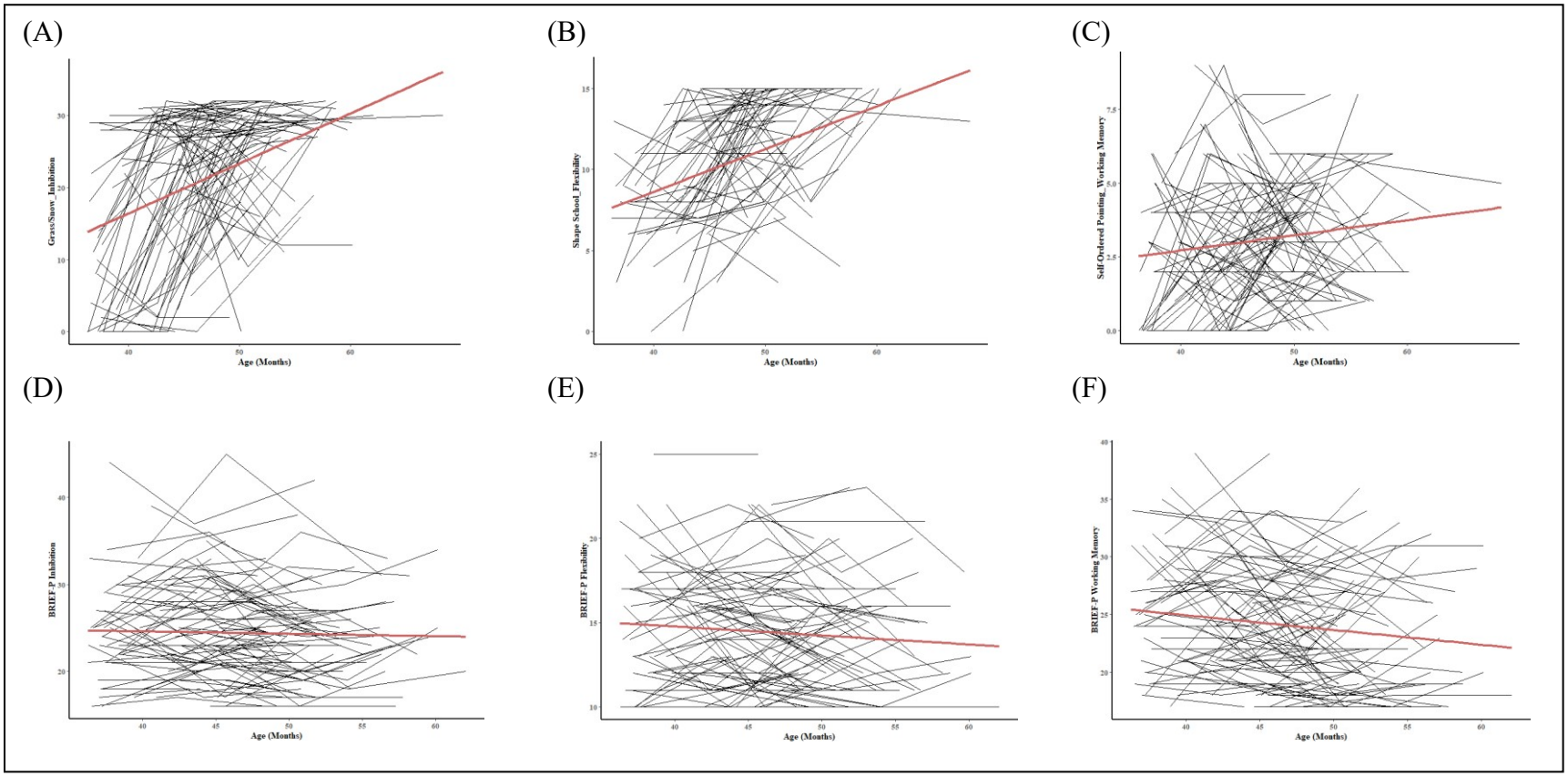
Demographic characteristics of participants

	Time 1			Time 2			Time 3		
Variable	*n*	*Mean (SD)*	Range	*n*	*Mean (SD)*	*Range*	*n*	*Mean (SD)*	*Range*
Age (months)	101	40.50 (3.23)	36.29-47.87	86	46.77 (3.47)	42.13-54.07	75	52.79 (3.59)	47.84-62.06
Sex (% girls)		48.51	-		50	-		49.33	-
Family Income					$100000- 150000[a]				
%Parental Postsecondary Education					92%				

[a] mode

Figure 1

Trajectories of EF components measured by PB tasks and BRIEF-P scales



Note. Individual raw score trajectories (black) and linear group trendlines (red) are plotted: (A) Grass/Snow task raw scores; (B) Shape School task raw scores; (C) Self-ordered Pointing task raw scores; (D) BRIEF-P Inhibit scale raw scores; (E) BRIEF-P Shift scale raw scores; (F) BRIEF-P Working Memory scale raw scores. BRIEF-P = Behaviour Rating Inventory of Executive Function-Preschool Version.

Table 2

Descriptive statistics and correlations for study variables

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Age1	101	40.50	3.23									
2. Age2	86	46.77	3.47	.98**								
3. Age3	75	52.79	3.59	.94**	.96**							
4. INH1	101	24.14	5.37	-.07	-.12	-.00						
5. INH2	86	24.78	5.84	-.19	-.20	-.06	.67**					
6. INH3	69	24.36	5.30	-.06	-.05	.04	.67**	.79**				
7. SHIFT1	101	14.73	3.62	-.03	-.06	-.09	.15	.06	.02			
8. SHIFT2	86	14.59	3.57	-.01	-.01	.02	.19	.43**	.23	.68**		
9. SHIFT3	69	13.86	3.10	-.02	-.02	-.02	.23	.31**	.33**	.62**	.75**	
10. WM1	101	24.37	5.51	-.21*	-.29**	-.23*	.70**	.41**	.44**	.23*	.15	.21
11. WM2	86	24.14	5.26	-.34**	-.33**	-.23*	.41**	.68**	.53**	.22*	.37**	.25*
12. WM3	69	23.96	5.05	-.14	-.14	-.08	.49**	.53**	.70**	.17	.25*	.30*
13. G/S1	89	16.18	10.96	.36**	.41**	.40**	-.20[a]	-.07	-.11	.03	.08	.02
14. G/S2	82	22.51	9.93	.23*	.23*	.17	-.16	-.13	-.26*	.13	.22*	.15
15. G/S3	73	25.42	6.53	.15	.13	.08	-.00	-.03	-.07	.14	.29*	.22
16. SS1	68	8.87	3.31	.30*	.35**	.48**	-.11	-.01	.00	-.16	-.04	-.01
17. SS2	72	10.76	3.23	.18	.20	.21	-.04	.02	-.11	-.04	.03	-.01
18. SS3	73	12.11	3.26	.31**	.31**	.24*	-.13	-.22	-.24	.12	.11	.09
19. SOP1	93	2.85	2.20	.28**	.30**	.28*	-.14	-.18	-.21	-.14	-.12	-.09
20. SOP2	79	3.16	2.05	.13	.11	.10	-.04	-.13	-.19	.02	-.02	.04
21. SOP3	73	3.21	2.05	.05	.08	.03	-.04	-.01	-.11	-.05	.21	.14

Variable	10	11	12	13	14	15	16	17	18	19	20	21
1. Age1												
2. Age2												
3. Age3												
4. INH1												
5. INH2												
6. INH3												
7. SHIFT1												
8. SHIFT2												
9. SHIFT3												
10. WM1												
11. WM2	.59**											
12. WM3	.62**	.68**										
13. G/S1	-.33**	-.14	-.25									
14. G/S2	-.23*	-.20	-.19	.42**								
15. G/S3	-.16	-.17	-.06	.22	.42**							
16. SS1	-.25*	-.01	.11	.39**	.08	.19						
17. SS2	-.01	.06	-.03	.28*	.20	.16	.30*					
18. SS3	-.24*	-.21	-.14	.22	.38**	.49**	.53**	.32*				
19. SOP1	-.20[a]	-.21	-.21	.16	.09	.04	.13	.13	.16			
20. SOP2	-.05	-.07	-.22	.14	.17	.26*	.27*	.14	.40**	.36**		
21. SOP3	-.14	-.08	-.23[a]	.19	.10	.22	.22	.26*	.21	.27*	.28*	

Note. *M* and *SD* used to represent mean and standard deviation, respectively. INH = BRIEF-P Inhibit scale; SHIFT = BRIEF-P Shift scale; WM = BRIEF-P Working Memory scale; G/S = Grass/Snow task; SS = Shape School task; SOP = Self-ordered Pointing task. Number on the right side of the measure indicates the time of assessment. Highlighted rectangles indicate direct correlations between performance-based measures and corresponding BRIEF-P scales.

[a] marginal significance (p = .056-.063). * $p < .05$. ** $p < .01$. *** $p < .001$.

Latent Variable Analysis

Prior to fitting the initial longitudinal CFA model, I specified an alternative longitudinal null model (i.e., no common factors, means and variances of the same indicator constrained to equivalence over time), χ^2 (177) = 922.466, $p < .001$. A longitudinal CFA model was then estimated, where all three PB measures loaded onto the latent factor PB, and BRIEF-P scales onto the latent factor BRIEF, testing configural invariance. This initial CFA model indicated a good model fit[3], χ^2 (102) = 105.189, $p = .395$, CFI = .996, TLI = .993, RMSEA = .018, and had statistically significant factor loadings for all indicators on corresponding factors except SOP at Time 1 ($\lambda = .25$, $p = .066$) and Time 3 ($\lambda = .23$, *ns;* see Table 3 for all loadings and Figure 2 for a visual representation).

3. The initial longitudinal measurement model converged with good model fit, but yielded a non-positive definitelatent covariance matrix ($r = 1.01$, $p < .001$) and residual covariance matrix. Following the guidelines provided by Kolenikov and Bollen (2012), such errors are likely due to the small and unequal sample sizes used in the current study rather than model misspecifications and thus no additional constraints were made.

Table 3

Factor loadings and intercepts from longitudinal measurement model

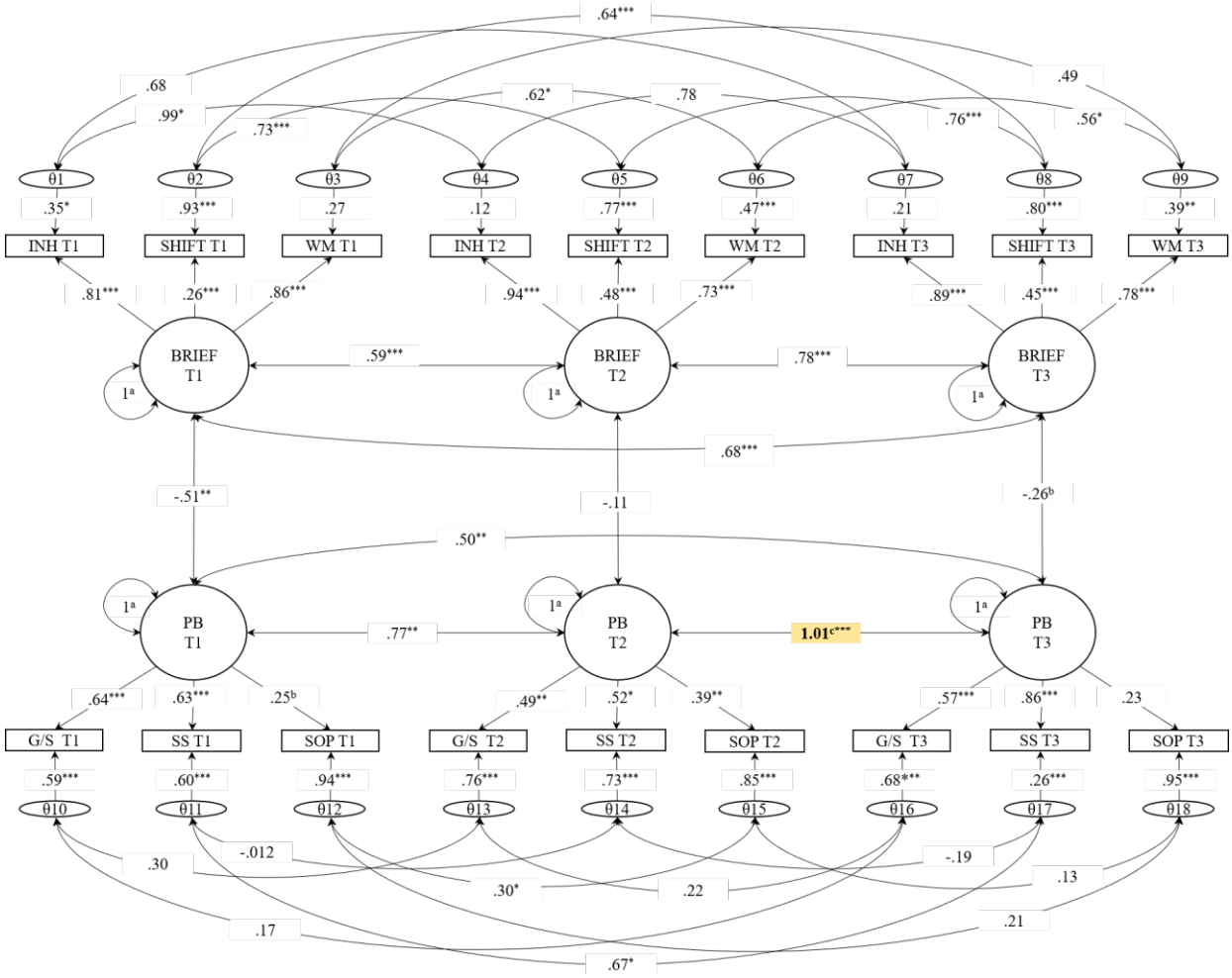
Latent Variable Indicator	Standardized Loading	Raw-Metric Loading (SE)	Standardized Intercepts	Raw-Metric Intercepts (SE)
Time 1 PB				
G/S	.64***	7.03 (1.34)	1.41***	15.52 (1.13)
SS	.63***	2.13 (.54)	2.55***	8.55 (.37)
SOP	.25[a]	.56 (.30)	1.27***	2.79 (.22)
Time 2 PB				
G/S	.49***	4.91 (1.66)	2.21***	22.14 (1.12)
SS	.52*	1.71 (.78)	3.22***	10.50 (.40)
SOP	.39**	.77 (.29)	1.54***	3.09 (.22)
Time 3 PB				
G/S	.57***	3.67 (.88)	3.94***	25.38 (.75)
SS	.86***	2.77 (.46)	3.72***	11.97 (.37)
SOP	.23	.46 (.31)	1.59***	3.21 (.24)
Time 1 BRIEF				
INH	.81***	4.39 (.58)	.45***	24.14 (.53)
SHIFT	.86***	.95 (.27)	4.03***	14.73 (.36)
WM	.86***	4.59 (.61)	4.55***	24.37 (.55)
Time 2 BRIEF				
INH	.94***	5.35 (.61)	4.31***	24.52 (.60)
SHIFT	.86***	1.70 (.33)	4.06***	14.48 (.37)
WM	.73***	3.97 (.50)	4.44***	24.09 (.56)
Time 3 BRIEF				
INH	.89***	4.55 (.55)	4.70***	24.01 (.56)
SHIFT	.45***	1.44 (.30)	4.36***	13.93 (.34)
WM	.78***	3.99 (.56)	4.69***	23.91 (.58)

[a] marginal significance ($p = .066$)

* $p < .05$. ** $p < .01$. *** $p < .001$.

Figure 2

Longitudinal measurement model



Note. Standardized values are shown. Circles represent latent factors and rectangles represent observed indicators. Ovals represent residual variances of indicators (θ). Double-headed arrows indicate correlations. Single-headed arrows from factors to indicators represent factor loadings (λ). [a] Fixed factor scaling method was used, constraining latent variances to 1. [b] Marginal significance (p = .64-.66). [c] Heywood case likely due to the small and unequal sample sizes.

* $p < .05$. ** $p < .01$. *** $p < .001$.

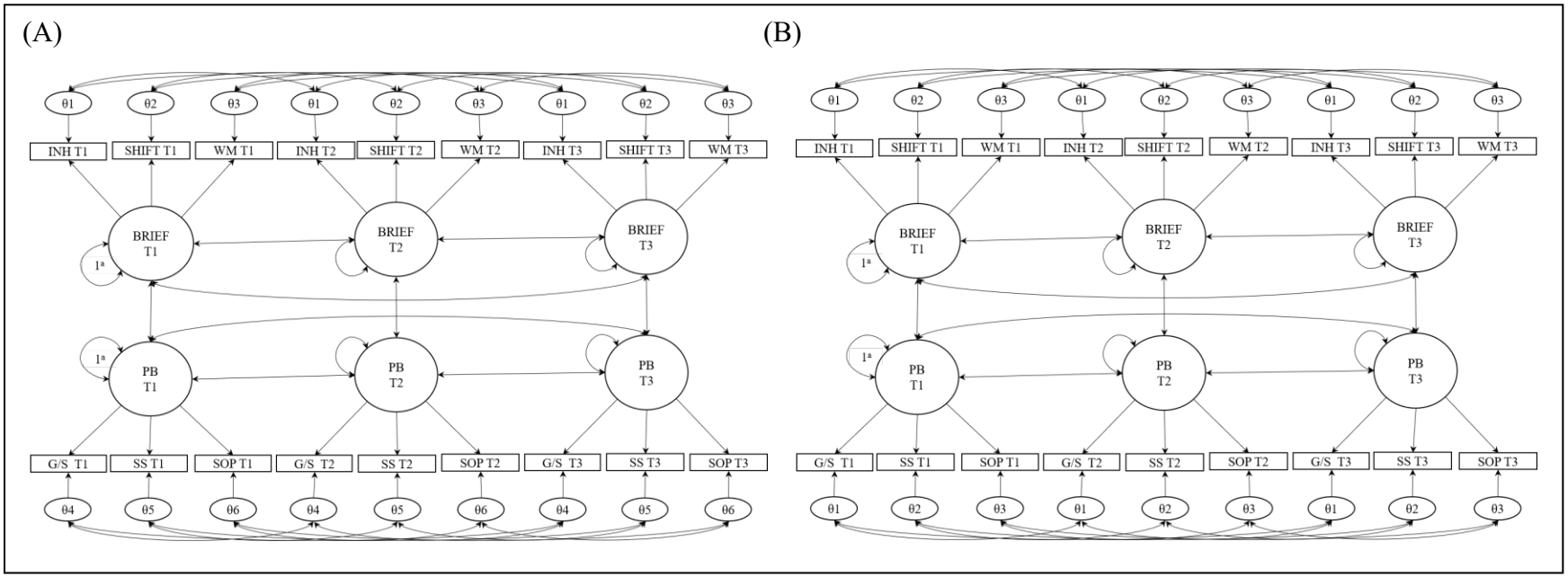
Do PB Measures and BRIEF-P Assess the Same Construct?

To examine whether PB measures and BRIEF-P tap into the same EF construct, I fit a model with all the indicators loading on to a single EF latent factor across time. This model significantly decreased the model fit, χ^2 (114) = 147.323, $p < .05$, ΔCFI = -.040, ΔTLI = -.06, ΔRMSEA = .036. This was consistent with weak to moderate cross-sectional latent factor correlations shown in the initial longitudinal CFA model (see Figure 2), only one of which was significant at Time 1 ($r = -.51$, $p < .01$). Composite reliability estimates were weak and not acceptable (ω = .13, .29, .33, for times 1 through 3, respectively).

To examine whether or not PB measures have greater task impurity than BRIEF-P, two separate measurement models were fit. First, I fit a model with residual variances of manifest variables (θ) constrained to equivalence across time *within* each factor. Then I fit a model with θ constrained to equivalence over time *across* both factors (see Figure 3 for visual representations of these models). While the equivalence across model had an acceptable model fit, χ^2 (117) = 160.077, $p < .01$, CFI/TLI = .952/.913, RMSEA= .060), the loss of model fit based on the multiple criteria was not trivial, suggesting that the equivalence within factor model was a better fitting model (see Table 5 for changes in model fit indices). On average, the amount of residual variability in the PB measures were higher (θ = .62) than in BRIEF-P scales (θ = .49). At the level of indicators, SOP and Shift scale had the highest level of unexplained variances (see Table 4 for estimates)

Figure 3

Visual representation of models with residual variances of indicators constrained to equivalence over time within and across factors



Note. Graphic representation of measurement models with residual variances of indicators (θ) constrained to equivalence over time: (A) within each latent factor and (B) across factors. Same drawing conventions as in previous figures.

[a] Fixed factor scale setting method was used.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4

Unexplained residual variances of indicators (equivalence within each factor)

Latent Variable Indicator	Std. Residual Variances (θ)[a]
Time 1 PB	
G/S	.06
SS	.80
SOP	.99
Time 2 PB	
G/S	.07
SS	.87
SOP	.98
Time 3 PB	
G/S	.14
SS	.66
SOP	.99
Time 1 BRIEF	
INH	.34
SHIFT	.91
WM	.36
Time 2 BRIEF	
INH	.27
SHIFT	.77
WM	.35
Time 3 BRIEF	
INH	.28
SHIFT	.80
WM	.31
Average across times 1-3	
PB	.62
G/S	.09
SS	.77
SOP	.98
BRIEF	.49
INH	.30
SHIFT	.82
WM	.34

Note. [a] All residual variances were significant (p <.001)

Longitudinal Measurement Invariance

The initial test of weak invariance (i.e., equivalent loadings across time) worsened the model fit (ΔCFI = -.012, ΔTLI = -.019, ΔRMSEA =.015). Modification indices suggested that G/S and SS loaded significantly different at Time 3. Freely estimating factor loadings of G/S and SS at Time 3 resulted in a model that supported partial weak variance with good model fit, χ^2 (108) = 116.174, ΔCFI = -.012, ΔTLI = -.011, ΔRMSEA =.015. Equating indicator intercepts across time points further supported partial strong invariance, χ^2 (116) = 130.661,ΔCFI = -.009, ΔTLI = -.012, ΔRMSEA =.008, rendering latent mean comparison tenable.

Table 5

Model fit statistics for the tests of longitudinal model invariance, residual variances, and one-factor model

Model	χ^{2a} (*df*)	CFI/ TLI	RMSEA	RMSEA 90%CI	$\Delta\chi^2$	*Δdf*	*p*	ΔCFI/ ΔTLI	ΔRMSEA
Alternative null model	922.466 (177)***	-	-		-	-	-	-	-
			Measurement model estimates						
Longitudinal CFA	105.189 (102)	.996/.993	.018	.000;.056	-	-	-	-	-
Weak invariance	122.290(110)	.984/.973	.033	.000;.063	17.318	8	.027	-.012/-.019	.015
Partial weak invariance	116.174(108)	.989/.982	.027	.000;.060	11.450	6	.077[b]	-.007/-.011	.009
Strong invariance	130.661 (116)	.980/.970	.035	.000;.064	14.228	8	.076	-.009/-.012	.008
Means equivalence	163.961 (118)**	.938/.908	.062	.036;.084	28.507	2	<.001	-.042/-.062	.027
BRIEF equivalence	131.410 (117)	.981/.971	.035	.000;.063	.972	1	.324	.001/.001	.000
PB equivalence	164.861 (117)**	.936/.903	.064	.038;.086	2397.692	1	<.001	-.045/-.067	.029
			Residual variances/One-factor model estimates						
θ equivalence within	149.944(114)*	.952/.925	.056	.026;.079	-	-	-	-	-
θ equivalence across	160.077 (117)**	.942/.913	.060	.034;.083	9.679	3	.022	-.010/-.013	.004
1- factor model	147.323(114)*	.955/.931	.054	.022;.078	37.752	12	<.001	-.040/-.06	.036

Note. [a] Models were estimated using robust maximum likelihood and thus χ^2 statistics reported here cannot be directly compared. χ^2 difference tests were robust difference test based on Yuan-Bentler (1996) and Satorra-Bentler (2010) method. [b] α/2 since two loadingswere freely estimated based on the modification indices.

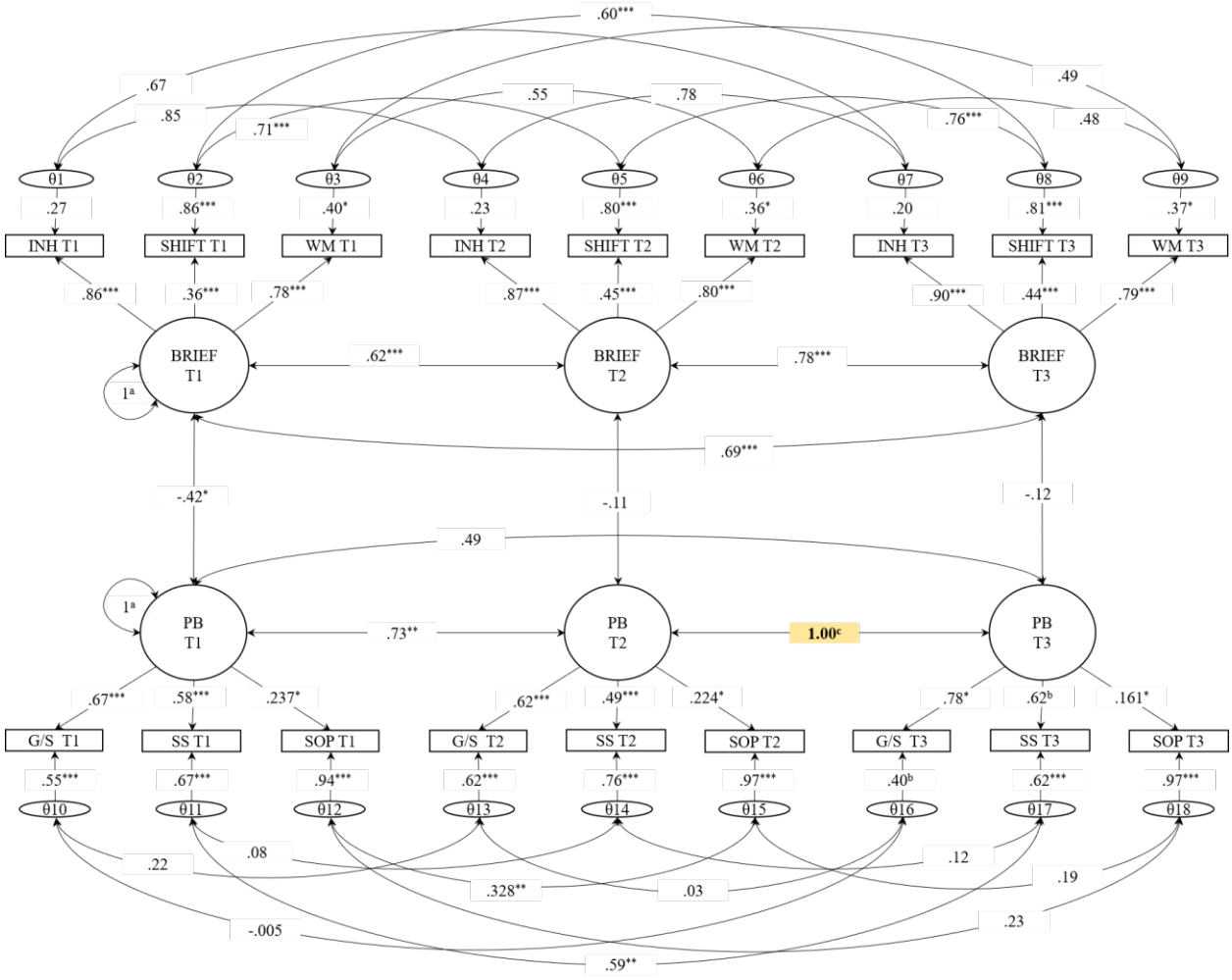
*$p < .05$. **$p < .01$. ***$p < .001$.

Do Children's EF Change over Time? A Latent Means Comparison

To test whether children's EF changed over time, regardless of the method of assessment, a latent means comparison model was fit with both PB and BRIEF latent means fixed to equivalence across all three time points. This model was significantly worse than the partial strong invariance model, χ^2 (118) = 163.961, $p < .01$, ΔCFI = -.042, ΔTLI = -.062, ΔRMSEA = .027, suggesting that the latent means changed over time. To test which (or whether both) of the PB and BRIEF latent means changed over time, two additional latent means models were fit, with respective latent means fixed to equivalence across time. The model with PB latent means freely estimated improved the model fit compared to the partial strong invariance model, χ^2 (117) = 131.410, $p = .171$, ΔCFI = .001, ΔTLI = .001, ΔRMSEA = .002, while the model with BRIEF latent means freely estimated had a significantly poorer fit (see Table 5 for fit indices). This indicated that children's EF assessed by PB measures significantly changed across time with constant latent mean when measured by BRIEF-P. The G/S task (λ = .62 -.78) and Inhibit scale (λ = .86 - .90) were the strongest indicator across time for PB and BRIEF factors, respectively. In contrast, SOP was the weakest loading indicator for PB with its loadings less than .35 across time points (λ = .16-.24), while the Shift scale was the weakest for the BRIEF factor (λ = .37-.45; see Table 6).

Figure 4

Latent means model with PB latent means freely estimated



Note. Standardized values are shown with same drawing conventions as previous figure. [a] Fixedfactor scaling method used with subsequent latent variances freely estimated. [b] Marginal significance ($p < .065$). [c] Correlation close to $r = 1.00$, potentially suggesting collinearity in PB latent factors at Times 2 and 3 although it was not significant ($p = .073$). * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 6

Factor loadings and intercepts from the latent means CFA model with PB latent means freely estimated

Latent Variable Indicator	Standardized Loading	Raw-Metric Loading (SE)	Standardized Intercepts	Raw-Metric Intercepts (SE)
Time 1 PB				
G/S	.67***	7.25 (1.27)	1.49***	16.05 (1.24)
SS	.58***	1.94 (.37)	2.56***	8.60 (.36)
SOP	.24*	.52 (.21)	1.22***	2.69 (.19)
Time 2 PB				
G/S	.62***	equated (T1)	1.61***	equated (T1)
SS	.49***	equated (T1)	2.56***	equated (T1)
SOP	.22*	equated (T1)	1.36***	equated (T1)
Time 3 PB				
G/S	.78*	8.36 (4.19)	2.42***	equated (T1)
SS	.62[a]	3.17 (1.66)	2.73***	equated (T1)
SOP	.161*	equated (T1)	1.34***	equated (T1)
Time 1 BRIEF				
INH	.86***	4.66 (.48)	4.17***	22.73 (.21)
SHIFT	.37***	4.05 (.81)	3.63***	13.87 (.31)
WM	.78***	1.41 (.23)	4.39***	22.87 (.25)
Time 2 BRIEF				
INH	.88***	equated (T1)	3.92***	equated (T1)
SHIFT	.45***	equated (T1)	4.01***	equated (T1)
WM	.80***	equated (T1)	4.15***	equated (T1)
Time 3 BRIEF				
INH	.90***	equated (T1)	4.38***	equated (T1)
SHIFT	.44***	equated (T1)	4.31***	equated (T1)
WM	.79***	equated (T1)	4.49***	equated (T1)

Note. [a] Marginal significance ($p = .056$).

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7

Means, variances, and correlations for latent means CFA model with PB latent means freely estimated

	1	2	3	4	5	6	Std. Mean
1. PB1	1.00						0.00
2. PB2	.73**	.72*					1.02***
3. PB3	.49	**1.00**[a]	.38				1.78[b]
4. BRIEF1	-.42*	-.25	-.12	1.00			.31**
5. BRIEF2	-.14	-.18	-.16	.62***	1.90***		.28**
6. BRIEF3	-.16	-.36	-.21	.69***	.78***	1.00***	.31**

Note. Values shown on diagonal are variances with correlations below diagonal. [a] Correlation close to $r = 1.00$, potentially suggesting collinearity in PB latent factors at times 2 and 3 but itwas not significant ($p = .073$). [b] Marginal significance ($p = .052$).

* $p < .05$. ** $p < .01$. *** $p < .001$.

Discussion

The purpose of this study was to examine the longitudinal relationship between performance-based (PB) and parent-rating (BRIEF-P) measures in typically developing preschoolers using a latent variable analytic approach. While these two types of measures are purported to capture the same construct, that is, EF, previous studies have shown that these two types of measures are largely unassociated. The current study attempted to further clarify the relationship between these two types of measures by examining (a) the longitudinal association between PB measures and corresponding BRIEF-P scales at both measurement and structural levels; (b) whether these measure types assess the same respective latent variables across time; (c) whether there are differences in the extent to which these measures are sensitive to children's EF development. Results suggested that there were both similarities and differences between these two types of measures. In the following sections, I attempt to position my interpretation of these results in thecontext of existing literature.

Cross-sectional and Longitudinal Correlations

Consistent with previous findings, individual PB measures were not associated with corresponding BRIEF-P scales across time. There was one exception to this lack of correlations in that the BRIEF-P WM scale was associated with G/S and SS tasks at the initial assessment, but not with the SOP task (see also Miranda et al., 2015, for similar results). Reviewing the items on the WM scale may explain these associations (or a lack thereof). WM is generally conceptualized as having two systems, the storage and central executive systems (also referred to as simple and complex WM, respectively; Baddeley & Hitch,1974; Garon et al., 2008). Yet, the items on the WM scale (e.g., "Has trouble remembering something, even after a brief period of time", "Cannot stay on the same topic when talking") seem to reflect primarily the storage

component of WM (vs. updating or combination of the two). Similarly, both G/S and SS tasks require children to hold the rules in mind (e.g., if told “grass”, point to white card; if with glasses, name shapes instead of colors) without having to update these rules to complete the tasks. In contrast, the SOP task is designed to assess both storage and updating components, where children have to remember the pictures they have previously pointed to and monitor their choices as they complete each trial. This finding is consistent with the previous study where the WM scale was associated with only simple, but not complex WM task (Garon et al., 2016), suggesting that the BRIEF-P WM scale may be a measure of information storage rather than updating (see also Loe et al., 2015).

Different patterns of within-measure-type interrelations emerged between PB measures and BRIEF-P scales. BRIEF-P scales were moderately correlated with each other across time. Specifically, the BRIEF-P Inhibit and WM scales were most strongly correlated at each time point. In contrast, the Shift scale had lower correlations with both the Inhibit and WM scales, the associations of which were attenuated compared to those reported in the technical manual for the BRIEF-P (Gioia et al., 2003). One of the potential reasons may be that the parental norms and expectations differ across the three scales. For instance, overall, the items on the Shift scale appear to tap into how children react to transitioning from one context to another (e.g., “Becomes upset with new situations”, “Has trouble adjusting to new people”). The items on Inhibit scale captures children’s behavioural regulation (e.g., “Is impulsive”, “Acts too wild or out of control”). WM scale represents children’s capacity to hold information in mind to complete a task (e.g., “Repeats the same mistakes over and over”, “Has trouble finishing tasks”). As such, there seems to be differential valence associated with each scale for parents such that the difficulties with the skills captured on the Inhibit and WM scales are more salient for parents

than those on the Shift scale. Alternatively, the attenuated associations may reflect that the items on the Shift scale are less well-defined and more heterogeneous compared to those on the Inhibit and WM scales. Consistent with this interpretation, the Shift scale had the lowest internal consistency on average among the three scales.

Interestingly, there were inconsistent correlations among the three PB tasks across time such that only G/S and SS tasks were moderately correlated at Times 1 and 3, while SOP was not correlated with either task across time. One of the challenges in assessing preschoolers' EF is that, given the rapid development of EF during this period, the same PB measures may assess different EF components and/or impose varying degrees of challenges across time (Barkley, 2012; Carlson et al., 2016). For instance, in the current study, while children's performance on all three tasks was significantly associated with children's age at first assessment, only G/S and SS tasks were associated with age at Times 2 and 3, respectively. This suggests that these tasks may be differentially sensitive to children's EF development across time. Alternatively, these inconsistent intercorrelations may be simply due to the similarities and differences in the nature of these tasks. For instance, both G/S and SS tasks included coloured stimuli and required some forms of interaction with examiners (e.g., pointing to either green or white card in response to the examiner's verbal prompts; naming colours or shapes in response to the stimuli pointed by examiners). In contrast, the SOP task included greyscale stimuli, and it was neither interactive (e.g., responses were *self*-ordered) nor required verbal responses (e.g., children were explicitly instructed to point without saying anything), which may have not been as appealing or motivating as the other two tasks for the current sample.

Willoughby and Blair (2016) suggested that these inconsistent intercorrelations among PB measures may reflect the lack of concordance between how the construct of EF is

conceptualized and measured. Specifically, these authors suggested that there are two overarching ways of conceptualizing EF with implications for its measurements. Broadly defined, the term EF refers to processes involved in the top-down control of cognition, emotion, and behaviours, and these control processes are often associated with the prefrontal cortex. This conceptualization assumes that the processes underlying this construct (e.g., inhibition, WM, flexibility) are interdependent, and such an assumption is translated to the common use of CFA in the literature, where PB tasks are modeled as reflective indicators whose variances (and thus individual differences in EF) are (assumed to be) accounted for by the common latent construct EF. Notwithstanding this assumption, however, previous studies have shown low to modest associations among PB tasks as in the current study (e.g., Willoughby et al., 2014). As such, Willoughby and Blair (2016) proposed an alternative conceptualization of EF, representing EF as emerging from individual differences in general abilities (e.g., attentional capacity and intelligence; see also Blair & Willoughby, 2013). In analytic terms, this alternative perspective corresponds to modelling PB tasks as formative indicators whose variances account for individual differences in EF. Although a broader discussion of the issue of whether reflective or formative measurement of EF is a better approach to capture EF is beyond the scope of the current study, the findings of this study suggest a need for a more refined conceptualization of EF (see later sections for further discussions).

Assessment of Executive Function Development

The current study found that children's EF improved over time, but only when measured by PB measures (see also Figure 1) and that the two latent variables, PB and BRIEF-P, were associated only at Time 1. These findings suggest that PB measures may be more sensitive to age-related differences (i.e., EF development) than BRIEF-P scales, consistent with previous

findings (e.g., Garon et al., 2016; Mahone & Hoffman, 2007). Likewise, the current study supported partial weak longitudinal invariance indicating that the relationship between the PB latent variable and two of its manifest variables (i.e., G/S and SS) changed over time, implying that the contribution of different components of EF to the latent structure changes during the preschool period, as suggested in previous studies (e.g., Garon et al., 2008; Miller et al., 2012; Wiebe et al.,2011). This was evident only in PB tasks, suggesting that PB measures may be more sensitive than BRIEF-P scales in detecting these changes. One notable finding was that the PB measures and BRIEF-P scales had both commonalities and differences in the patterns of factor loadings across time. Specifically, both PB and BRIEF-P had inhibition as the strongest loading indicator (i.e., G/S task and Inhibit scale, respectively). However, while both had an indicator that loaded at either inconsequential or minimally meaningful extent, WM indicator was the least loading for PB, whereas flexibility indicator was the least loading for BRIEF-P. This further suggests that these two types of measures differ in the extent to which they are sensitive to capture different components of EF.

Rating measures, including the BRIEF-P, are often developed to capture behavioural manifestations of children's EF (Roth et al., 2014). However, these ratings are usually in terms of difficulties and impairments (as in the BRIEF-P), and thus have been extensively used in clinical assessments, often to distinguish clinical groups (e.g., ADHD; Ezpeleta et al., 2015) from typically developing children. This, in turn, suggests that rating measures may be more sensitive for measuring EF in atypically developing children in that they focus on dysfunctions, rather than the full spectrum of EF, and thus may be less useful for measuring EF in typically developing children, as used in the current study (Anderson et al., 2002; McAuley et al., 2010).

Do They Assess the Same Construct?

When BRIEF-P scales and corresponding PB measures were loaded onto one single factor across time, it significantly worsened the model fit together with unacceptable reliability coefficients. This finding suggests that PB tasks and corresponding BRIEF-P scales do not measure the same construct, or at least, the same aspects of EF.

Traditionally, PB measures have been criticized for limited ecological validity because of their abstract and decontextualized nature of assessment, often limited to assessing circumscribed (or narrow) cognitive control processes, while neglecting to account for real-world context that often involves emotional control processes (Barkley, 2012). This emotional and motivational component of EF is reflected in studies examining the distinction between "hot" and "cool" EF processes (e.g., Zelazo & Carlson, 2012; Zelazo & Müller, 2002). Hot EF processes are the skills needed in emotionally or motivationally charged situations, whereas cool EF processes draw on those elicited in emotionally neutral and decontextualized situations. From this perspective, PB tasks used in the current study were cool EF tasks, designed to capture specific cognitive components of EF (i.e., inhibition, WM, and flexibility). In contrast, the rationale for the development of initial BRIEF was to acknowledge that EF is not exclusive to cognitive control processes but also includes behavioral and emotional control processes (Isquith et al., 2005). This hot vs. cool distinction is also evident at the level of neural substrates, where performances on hot EF tasks (e.g., Iowa Gambling Task[4]; Bechara et al., 1994, 1997) and cool EF tasks (e.g., DCCS; Zelazo, 2006) have been differentially associated with orbitofrontal and dorsolateral prefrontal cortex, respectively. Thus, it is perhaps not surprising that these two types of measures are not associated with each other.

4. Bechara et al. (1994, 1997) provided evidence for the contribution of orbitofrontal cortex to the hot EF processesusing adult samples. Kerr and Zelazo (2004) used Children's Gambling Task (a child version of Iowa GamblingTask) to examine the development of affective decision-making and found that 4-year-olds performed better than3-year-olds. Based on these findings, Kerr and Zelazo (2004) inferred that these age differences in performances correspond to the age-related development of orbitofrontal cortex, providing indirect support for Bechara et al.'s findings.

The finding that PB measures and BRIEF-P scales were not represented better as a single factor is in line with Toplak et al.'s (2013) distinction between typical and optimal performance, further reflected in the way that these measures were administered in the current study. For instance, given the highly structured and controlled nature of PB tasks, where children are provided with a set of explicit instructions, Toplak et al. suggested that these tasks tap into how efficiently children can process the given information and achieve the goals specified by examiners. In contrast, on BRIEF-P, while parents are explicitly instructed to estimate the frequency of children's day-to-day EFs over the past 6 months, they are not prescribed to any specific everyday situations to reflect on. Instead, the interpretation of the items on BRIEF-P and the specific situations to reflect on are at the discretion of the parents, rendering room for subjectivity and biases that are often inherent in rating measures (Denckla, 2002). This account may explain the current finding where BRIEF-P scales, compared to PB tasks, had not only significant but also higher intercorrelations, potentially indicating the presence of the halo effect (Thorndike, 1920). Parents may also be estimating these everyday EFs against their children's own baseline in context, taking into account various day-to-day situational and environmental factors (e.g., stress, fatigue, and attention; Mattson et al., 2020). In a similar vein, the parent ratings may also be anchored on certain components of EF that the parents perceive as the most important or relevant for their children. This interpretation is consistent with the previous finding where children with better parent-rated WM *relative* to parent-rated inhibition performed better on a WM task (Garon et al., 2016).

On the other hand, to reiterate, PB tasks are designed to assess specific EF components, and children's performance on these tasks are often considered as *absolute* indices of corresponding components. While it is often assumed that the aforementioned extraneous

contextual factors can affect children's task performance to a varying extent depending on the child and/or the task itself, these assumptions often go untested. Related to this, Mattson et al. (2020) recently found that children's parent-rated and PB EFs were significantly associated when children's performance was operationalized in terms of split-half task performance to reflect within-task variability (i.e., the difference in performance between first and second halves of the task). These authors proposed that the lack of correlations between PB and rating measures may be in part attributable to the common use of summary scores to reflect children's task performance, thus neglecting to account for the individual variability that may be at play.

Implications

Despite the considerable amount of empirical and theoretical attention devoted to the construct of EF, there is currently no universally accepted operational definition of EF including a lack of consensus on the structure of EF across the lifespan, which skills are subsumed by this construct, and most importantly, how these skills should be measured (Carlson et al., 2016; Müller & Kerns, 2015; see also Nigg, 2017, for a discussion on the range of various terms associated with EF). For instance, the construct of EF and its components are often referred to with interchangeable names (e.g., executive function vs. executive functioning/control; inhibition vs. inhibitory control; [set-]shifting vs. flexibility). Furthermore, several factor analytic studies have previously examined the structure of EF, reporting mixed findings with respect to the dimensionality of EF in preschoolers, ranging from unitary (e.g., Hughes et al., 2010; Wiebe et al., 2008, 2011) to three-factor structure (e.g., Howard et al., 2015), suggesting that the structure of EF changes across childhood, becoming more differentiated (see Bardikoff & Sabbagh, 2017; Karr et al., 2018, for a review).

A relatively recent systematic review of 106 studies on EF showed both convergence and

divergence in how these studies conceptualized and operationalized the construct of EF (Baggetta & Alexander, 2016). Perhaps what stood out the most from this review was not only that it identified a total of 109 different PB tasks used in these studies, but also that over 27% of these tasks were used either to assess multiple components or as an overall index of the construct EF itself (e.g., Stroop task and its variants being used to measure either inhibition, WM, attention, or overall EF; see Baggetta & Alexander, 2016, for more details). As such, the lack of correlations between PB tasks and BRIEF-P scales may reflect the discordance in how researchers conceptualize EF when developing and selecting different measures. For instance, the component inhibition is generally considered to consist of several subcomponents such as response inhibition and conflict inhibition (also known as interference control; Nigg, 2000). These distinctions are evident in previous neuroimaging findings where different neural substrates within prefrontal cortex regions were associated with different types of inhibition (see Friedman & Miyake, 2004; Nigg, 2000, for a brief overview). As such, PB tasks commonly used to assess these subtypes of inhibition often differ in nature (e.g., Go/No-go task for response inhibition, Day/Night task for conflict inhibition; Anderson & Reidy, 2012; Carlson, 2005; Montgomery & Koeltzow, 2010). In contrast, the items on the BRIEF-P Inhibit scale do not seem to differentiate these distinct subcomponents of inhibition. A similar case can be made for the previously discussed differences between the SOP task and the BRIEF-P WM scale. Thus, future studies using more integrated PB tasks and rating scales may find better associations between these two types of measures.

The current study conceptualized EF as having three main components (i.e., inhibition, WM, and flexibility); however, such a structure is not universally agreed upon. Indeed, there is a lack of coherent and unified theoretical grounds to assume that EF is best characterized by

these three components (Müller & Kerns, 2015). For instance, while the origins of the three-component representation of EF are often traced back to the seminal work of Miyake et al. (2000), the purpose of this work was not to demonstrate that EF is best represented by these three components, but rather, how these three components differentially contribute to EF (Doebel, 2020). Indeed, the initial selection of these three components was based on previous literature where these three components were the most frequently posited ones (Miyake et al., 2000; see also Miyake & Friedman, 2012, for an update on their initial unity vs. diversity model).

Along these lines, Doebel (2020) recently proposed that our current conceptualization of EF as an interplay of different separable, yet related, components, is not only as theoretically or empirically grounded as often assumed, but also, does not reflect the nature of EF in real-world settings:

> …instead of thinking of the development of EF as the emergence of separable components that can themselves be meaningfully separated from task-specific demands… think of it as the *development of skills in using control in the service of specific goals* [emphasis added]. Critically, specific goals activate mental content such as relevant knowledge, beliefs, values, norms, interests, and preferences that children acquire with development in a specific sociocultural context, shaping how they use control. (p. 945)

Accounting for the discrepancy between these two types of measures, Doebel (2020) suggested that PB measures assess children's *EF capacity* while rating measures assess children's *skills in using EF* in everyday situations. While acknowledging the value in using PB tasks that are targeted to measure specific EF components depends on one's research questions (see also Friedman & Banich, 2019), Doebel (2020) suggested that we should not ignore the evident need

for more ecologically valid measures in order to address how children use EF in everyday situations.

Taken together, PB measures and BRIEF-P scales seem to assess different aspects of EF, providing complementary perspectives on children's EF. The divergence between these two types of measures do not necessarily indicate that one is better than the other. Instead, it may provide important information about the contexts in which children find difficulties in using their EF. Thus, I concur with the overarching theme of existing literature that both are needed and that one should not exclusively rely on either when assessing children's EF.

Limitations and Recommendations for Future Studies

The current study has a few limitations that should be considered when interpreting the results. The first line of limitations pertains to the measures used in the current study. First, it is known that the choice and number of EF tasks included can affect CFA results, yielding different EF structures (Miller et al., 2012). However, as the current study was part of a larger study, I was limited to only one measure for each EF component per measure type and thus, I was unable to examine the structure of EF, which may have provided a more nuanced understanding of the similarities and differences between PB tasks and BRIEF-P. Second, PB tasks are known to differ in their measurement precision such that some may be better at differentiating children within high levels of performance, while others are better at differentiating those within low levels of performance (Willoughby et al., 2012). For instance, as presented in Figure 1, children's performances on G/S and SS tasks appear to reach the ceiling towards the end of the study, with slower rates of improvement on the SOP task. The SOP task also had the greatest amount of unexplained variance across time, suggesting that the use of this task may not have been appropriate for the current sample compared to the G/S and SS tasks.

It is noteworthy that in the current study, both SS and SOP tasks were not administered as in their original versions. In contrast to the original administration of the SOP task (Petrides & Milner, 1982), in the current study, once children made an error on a trial, they were directed to move on to the next trial to minimize potential fatigue in children. However, such a protocol restricted the outcome variable to the number of correct sets (vs. error percentage and perseverance rates) and thus may have limited the score variability that might have been observed otherwise. Similarly, the original SS task is administered in the format of a storybook and includes four conditions (Espy et al., 2006). It starts with the control condition, where children simply have to name colours. In the second condition, they are told to name colours but only for happy faces (i.e., inhibition condition). The third condition requires children to name colours but only for those without hats while naming shapes for those with hats (i.e., switching condition). The last condition combines inhibition and switching, where children need to follow the switching rule but only for those with happy faces. The current study only included the conditions that are comparable to the original control and shifting conditions and as such, it was not possible to obtain more in-depth information (e.g., intra-individual task performance, comparing children with better performance in shifting relative to inhibition conditions). Additionally, research shows that the age-related improvements on the SS task are best reflected in switching fluency (i.e., efficiency; Clark et al., 2013). However, in the current study, given that children were often observed to frequently stop and engage with examiners during the task, the use of efficiency score was deemed unreliable. Specifically, the efficiency scores would likely have been confounded by children's level of attention. As such, with the absence of a control measure, I was restricted to the number of correct items as the outcome variable, which may have contributed to the observed ceiling effect.

One of the limitations of BRIEF-P is that the items on these scales are often strongly associated with clinical symptoms, for example, overlapping with ADHD diagnostic criteria (Mahone & Hoffman, 2007; Miranda et al., 2015). Such limitations have led to the development of new rating measures. For example, Nilsen et al. (2017) recently developed the Ratings of Everyday EF, a parent-rating EF measure based on children's behaviours in specific, but everyday contexts (e.g., peer interaction and home environment). Similarly, Thorell and Nyberg (2008) developed the Childhood EF Inventory as a potential alternative rating measure to BRIEF-P with shorter scales and as few items confounding with ADHD symptoms as possible (see Thorell et al., 2010, for clinical utility). Furthermore, as previously noted, studies have shown discordance between parent- and teacher-reported EF. Thus, time and resources permitting, future studies should consider using a larger battery of PB tasks (e.g., National Institute of Health Toolbox Cognition Battery; Zelazo et al., 2013), together with multiple rating measures from multiple informants to obtain a more comprehensive understanding of preschoolers' EF and its assessment.

The second line of limitations pertains to the sample used in the study. First, not all children participated in all three time points, resulting in small and unequal sample sizes across time. Attrition is often inherent in longitudinal studies and there are several statistical techniques to account for attrition a priori and to better estimate missingness due to attrition (planned missing data designs; Little et al., 2015). While I used FIML to handle missingness in the data, it should be noted that I did not directly test the randomness of missingness, that is, whether the missingness was completely at random, at random, or not at random. The missingness in the current data may have been associated with auxiliary variables such as children's socioeconomic status, gender, and family dynamics (e.g., parental involvement and presence of siblings). This

should also be considered in view of the COVID-19 pandemic, which may have differentially affected families' motivation to continue participating in the current study. Thus, future studies should analyze the mechanisms of missingness, which may or may not have attenuated the potential association between children's PB and parent-rated EF.

The current sample may not be representative of children in BC and Canada, limiting the generalizability of the current findings. For instance, the percentage of parental post-secondary education (92%) and the median household income range ($100000-150000) were both higher than the national average (54% and $70336, respectively; Statistics Canada, 2016). In addition, the current sample was homogeneous in that children were typically developing three-year-olds, recruited from the same general geographic location, and mostly White. Considering BRIEF-P often has better clinical utility in identifying clinical groups than PB measures, such homogeneity of the current sample may have attenuated the findings. Thus, future studies should recruit more diverse samples. It is also noteworthy that while the three assessments were six months apart, BRIEF-P also prompts to reflect on the past six months, which may have limited its sensitivity to reflect improvements in children's EF. Lastly, given previous findings that PB and rating measures have differential predictive validity with regard to children's subsequent developmental outcomes (e.g., Schmitt et al., 2014), it is recommended that future studies examine the directionality between these two types of measures.

Conclusion

The preschool period is an important developmental phase during which EF undergoes significant changes, which, in turn, are associated with children's subsequent functioning. The current study aimed to examine the longitudinal relationship between performance-based measures and BRIEF-P that are commonly used to assess EF in preschoolers. Consistent with

previous studies, these two types of measures were not associated with each other, both concurrently and longitudinally. These findings were interpreted in the context of the existing literature, leading to the conclusion that these two types of measures assess different aspects of children's EF, and that there is a need for a better conceptualization of EF and its development in children.

References

Acar, I. H., Frohn, S., Prokasky, A., Molfese, V. J., & Bates, J. E. (2019). Examining the associations between performance based and ratings of focused attention in toddlers: Are we measuring the same constructs? *Infant and Child Development*, *28*(1), e2116. https://doi.org/10.1002/icd.2116

Achenbach, T. M., Ivanova, M. Y., & Rescorla, L. A. (2017). Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive Psychiatry*, *79*, 4–18. https://doi.org/10.1016/j.comppsych.2017.03.006

Allan, N. P., Hume, L. E., Allan, D. M., Farrington, A. L., & Lonigan, C. J. (2014). Relations between inhibitory control and the development of academic skills in preschool and kindergarten: A meta-analysis. *Developmental Psychology*, *50*(10), 2368–2379. https://doi.org/10.1037/a0037493

Anderson, V. A., Anderson, P., Northam, E., Jacobs, R., & Mikiewicz, O. (2002). Relationships between cognitive and behavioral measures of executive function in children with brain disease. *Child Neuropsychology*, *8*(4), 231–240. https://doi.org/10.1076/chin.8.4.231.13509

Anderson, P. J., & Reidy, N. (2012). Assessing executive function in preschoolers. *Neuropsychology Review*, *22*(4), 345–360. https://doi.org/10.1007/s11065-012-9220-3

Baggetta, P., & Alexander, P. A. (2016). Conceptualization and operationalization of executive function. *Mind, Brain, and Education*, *10*(1), 10–33. https://doi.org/10.1111/mbe.12100

Bardikoff, N., & Sabbagh, M. (2017). The differentiation of executive functioning across development: Insights from developmental cognitive neuroscience. In N. Budwig, E.

Turiel, & P. D. Zelazo (Eds.), *New perspectives on human development* (pp. 47–66). Cambridge University Press. https://doi.org/10.1017/CBO9781316282755.005

Barkley, R. A. (2012). *Executive functions: What they are, how they work, and why they evolved*. Guilford Press.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15. https://doi.org/10.1016/0010-0277(94)90018-3

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293–1295.

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*(6), 1641–1660.

Blair, C., Protzko, J., & Ursache, A. (2011). Self-regulation and early literacy. In S. B. Neuman& D. K. Dickinson (Eds.), *Handbook of early literacy research* (Vol. 3, pp. 20-35). Guilford Press.

Blair, C., & Willoughby, M. (2013). Rethinking executive functions: Commentary on "The contribution of executive function and social understanding to preschoolers' letter and math skills" by M.R. Miller, U. Müller, G.F. Giesbrecht, J.I.M. Carpendale, and K.A. Kerns. *Cognitive Development*, *28*(4), 350–353. https://doi.org/10.1016/j.cogdev.2013.06.001

Bünger, A., Urfer-Maurer, N., & Grob, A. (2021). Multimethod assessment of attention, executive functions, and motor skills in children with and without ADHD: Children's performance and parents' perceptions. *Journal of Attention Disorders, 25*(4), 596-606. https://doi.org/10.1177/1087054718824985

Camerota, M., Willoughby, M. T., Kuhn, L. J., & Blair, C. B. (2018). The Childhood Executive Functioning Inventory (CHEXI): Factor structure, measurement invariance, and correlates in US preschoolers. *Child Neuropsychology*, *24*(3), 322–337. https://doi.org/10.1080/09297049.2016.1247795

Campbell, S. B., Denham, S. A., Howarth, G. Z., Jones, S. M., Whittaker, J. V., Williford, A. P., Willoughby, M. T., Yudron, M., & Darling-Churchill, K. (2016). Commentary on the review of measures of early childhood social and emotional development: Conceptualization, critique, and recommendations. *Journal of Applied Developmental Psychology*, *45*, 19–41. https://doi.org/10.1016/j.appdev.2016.01.008

Caporaso, J. S., Boseovski, J. J., & Marcovitch, S. (2019). The individual contributions of three executive function components to preschool social competence. *Infant and Child Development*, *28*(4), e2132. https://doi.org/10.1002/icd.2132

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology*, *28*(2), 595–616. https://doi.org/10.1207/s15326942dn2802_3

Carlson, S. M., Faja, S., & Beck, D. M. (2016). Incorporating early development into the measurement of executive function: The need for a continuum of measures across development. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research* (p. 45–64). American Psychological Association. https://doi-org.ezproxy.library.uvic.ca/10.1037/14797-003

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053.

Chaytor, N., Schmitter-Edgecombe, M., & Burr, R. (2006). Improving the ecological validity of executive functioning assessment. *Archives of Clinical Neuropsychology*, *21*(3), 217–227. https://doi.org/10.1016/j.acn.2005.12.002

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, Y.-C., Hwang-Gu, S.-L., Ni, H.-C., Liang, S. H.-Y., Lin, H.-Y., Lin, C.-F., Tseng, Y.-H., & Gau, S. S.-F. (2017). Relationship between parenting stress and informant discrepancies on symptoms of ADHD/ODD and internalizing behaviors in preschool children. *PLoS ONE*, *12*(10), e0183467. https://doi.org/10.1371/journal.pone.0183467

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi-org.ezproxy.library.uvic.ca/10.1207/S15328007SEM0902_5

Clark, C. A. C., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental Psychology*, *46*(5), 1176–1191. https://doi.org/10.1037/a0019672

Craig, F., Margari, F., Legrottaglie, A. R., Palumbi, R., de Giambattista, C., & Margari, L. (2016). A review of executive function deficits in autism spectrum disorder and attention-deficit/hyperactivity disorder. *Neuropsychiatric Disease and Treatment*, *12*, 1191–1202. https://doi.org/10.2147/NDT.S104620

Daunhauer, L. A., Gerlach-McDonald, B., Will, E., & Fidler, D. J. (2017). Performance and ratings based measures of executive function in school-aged children with Down syndrome. *Developmental Neuropsychology*, *42*(6), 351–368.

https://doi.org/10.1080/87565641.2017.1360303

Davidson, F., Cherry, K., & Corkum, P. (2016). Validating the Behavior Rating Inventory of Executive Functioning for children with ADHD and their typically developing peers. *Applied Neuropsychology: Child*, *5*(2), 127–137. https://doi.org/10.1080/21622965.2015.1021957

Dekker, M. C., Ziermans, T. B., Spruijt, A. M., & Swaab, H. (2017). Cognitive, parent and teacher rating measures of executive functioning: Shared and unique influences on school achievement. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00048

Denckla, M. B. (2002). The Behavior Rating Inventory of Executive Function: Commentary. *Child Neuropsychology*, *8*(4), 304–306.

Diamond, A. (2016). Why improving and assessing executive functions early in life is critical. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research.* (pp. 11–43). American Psychological Association. https://doi.org/10.1037/14797-002

Doebel, S. (2020). Rethinking Executive Function and Its Development. *Perspectives on Psychological Science*, *15*(4), 942–956. https://doi.org/10.1177/1745691620904771

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004

Espy, K. A. (1997). The shape school: Assessing executive function in preschoolchildren. *Developmental Neuropsychology, 13*(4), 495-499. https://doi.org/10.1080/87565649709540690

Ezpeleta, L., & Granero, R. (2015). Executive functions in preschoolers with ADHD, ODD, and

comorbid ADHD-ODD: Evidence from ecological and performance-based measures. *Journal of Neuropsychology*, *9*(2), 258–270. https://doi.org/10.1111/jnp.12049

Ezpeleta, L., Granero, R., Penelo, E., de la Osa, N., & Domènech, J. M. (2015). Behavior Rating Inventory of Executive Functioning–Preschool (BRIEF-P) applied to teachers: Psychometric properties and usefulness for disruptive disorders in 3-year-old preschoolers. *Journal of Attention Disorders*, *19*(6), 476–488. https://doi.org/10.1177/1087054712466439

Faridi, N., Karama, S., Burgaleta, M., White, M. T., Evans, A. C., Fonov, V., Collins, D. L., & Waber, D. P. (2015). Neuroanatomical correlates of behavioral rating versus performance measures of working memory in typically developing children and adolescents. *Neuropsychology*, *29*(1), 82–91. https://doi.org/10.1037/neu0000079

Fay-Stammbach, T., & Hawes, D. J. (2019). Caregiver ratings and performance-based indices of executive function among preschoolers with and without maltreatment experience. *Child Neuropsychology*, *25*(6), 721–741. https://doi.org/10.1080/09297049.2018.1530344

Friedman, N. P., & Banich, M. T. (2019). Questionnaires and task-based measures assessdifferent aspects of self-regulation: Both are needed. *Proceedings of the National Academy of Sciences*, *116*(49), 24396–24397. https://doi.org/10.1073/pnas.1915315116

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Fuhs, M. W., & Day, J. D. (2011). Verbal ability and executive functioning development in preschoolers at head start. *Developmental Psychology*, *47*(2), 404–416. https://doi.org/10.1037/a0021065

Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology*, *50*(6), 1698–1709. https://doi.org/10.1037/a0036633

Gardiner, E., Hutchison, S. M., Müller, U., Kerns, K. A., & Iarocci, G. (2017). Assessment of executive function in young children with and without ASD using parent ratings and computerized tasks of executive function. *The Clinical Neuropsychologist*, *31*(8), 1283–1305. https://doi.org/10.1080/13854046.2017.1290139

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*(1), 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Garon, N. M., Piccinin, C., & Smith, I. M. (2016). Does the BRIEF-P predict specific executive function components in preschoolers? *Applied Neuropsychology: Child*, *5*(2), 110–118. https://doi.org/10.1080/21622965.2014.1002923

Gerst, E. H., Cirino, P. T., Fletcher, J. M., & Yoshida, H. (2017). Cognitive and behavioral rating measures of executive function as predictors of academic outcomes in children. *Child Neuropsychology*, *23*(4), 381–407. https://doi.org/10.1080/09297049.2015.1120860

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 312–7 years old on a Stroop- like Day-Night test. *Cognition*, *53*(2), 129–153. https://doi.org/10.1016/0010-0277(94)90068-X

Gioia, G. A., Espy, K. A., & Isquith, P. K. (2003). *Behavior Rating Inventory of Executive Function - Preschool Version*. Psychological Assessment Resources.

Gioia, G. A., Isquith, P. K., Retzlaff, P. D., & Espy, K. A. (2002). Confirmatory factor analysis of the Behavior Rating Inventory of Executive Function (BRIEF) in a clinical sample. *Child*

Neuropsychology, *8*(4), 249–257.

Gross, A. C., Deling, L. A., Wozniak, J. R., & Boys, C. J. (2015). Objective measures of executive functioning are highly discrepant with parent-report in fetal alcohol spectrum disorders. *Child Neuropsychology, 21*(4), 531–538. https://doi.org/10.1080/09297049.2014.911271

Howard, S. J., Okely, A. D., & Ellis, Y. G. (2015). Evaluation of a differentiation model of preschoolers' executive functions. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00285

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hughes, C., Ensor, R., Wilson, A., & Graham, A. (2010). Tracking executive function across the transition to school: A latent variable approach. *Developmental Neuropsychology*, *35*(1), 20–36. https://doi.org/10.1080/87565640903325691

Isquith, P. K., Crawford, J. S., Espy, K. A., & Gioia, G. A. (2005). Assessment of executive function in preschool-aged children. *Mental Retardation and Developmental Disabilities Research Reviews*, *11*(3), 209–215. https://doi.org/10.1002/mrdd.20075

Isquith, P. K., Roth, R. M., & Gioia, G. (2013). Contribution of rating scales to the assessment of executive functions. *Applied Neuropsychology: Child*, *2*(2), 125–132. https://doi.org/10.1080/21622965.2013.748389

Jacobson, L. A., Schneider, H., & Mahone, E. M. (2018). Preschool inhibitory control predicts ADHD group status and inhibitory weakness in school. *Archives of Clinical Neuropsychology*, *33*(8), 1006–1014. https://doi.org/10.1093/arclin/acx124

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*, *144*(11), 1147–1185. https://doi.org/10.1037/bul0000160

Kolenikov, S., & Bollen, K. A. (2012). Testing negative error variances: Is a heywood case a symptom of misspecification? *Sociological Methods & Research*, *41*(1), 124–167. https://doi.org/10.1177/0049124112442138

Liebermann, D., Giesbrecht, G. F., & Müller, U. (2007). Cognitive and emotional aspects of self-regulation in preschoolers. *Cognitive Development*, *22*(4), 511–529. https://doi.org/10.1016/j.cogdev.2007.08.005

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Little, T. D., Deboeck, P., & Wu, W. (2015). Longitudinal data analysis. In S. M. Kosslyn & R. A. Scott (Eds.), *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*. Wiley. https://doi.org/10.1002/9781118900772

Loe, I. M., Chatav, M., & Alduncin, N. (2015). Complementary assessments of executive function in preterm and full-term preschoolers. *Child Neuropsychology*, *21*(3), 331–353. https://doi.org/10.1080/09297049.2014.906568

Lonigan, C. J., Spiegel, J. A., Goodrich, J. M., Morris, B. M., Osborne, C. M., Lerner, M. D., & Phillips, B. M. (2017). Does preschool self-regulation predict later behavior problems in general or specific problem behaviors? *Journal of Abnormal Child Psychology*, *45*(8), 1491–1502. https://doi.org/10.1007/s10802-016-0260-7

Mahone, E. M., & Hoffman, J. (2007). Behavior ratings of executive function among preschoolers with ADHD. *The Clinical Neuropsychologist*, *21*(4), 569–586.

https://doi.org/10.1080/13854040600762724

Mahone, E. M., Martin, R., Kates, W. R., Hay, T., & Horská, A. (2009). Neuroimaging correlates of parent ratings of working memory in typically developing children. *Journal of the International Neuropsychological Society*, *15*(1), 31–41. https://doi.org/10.1017/S1355617708090164

Malloy, P., & Grace, J. (2005). A review of rating scales for measuring behavior change due to frontal systems damage: *Cognitive and Behavioral Neurology*, *18*(1), 18–27. https://doi.org/10.1097/01.wnn.0000152232.47901.88

Mattson, J. T., Thorne, J. C., & Kover, S. T. (2020). Relationship between task-based and parent report-based measures of attention and executive function in children with Fetal Alcohol Spectrum Disorders (FASD). *Journal of Pediatric Neuropsychology*, *6*(3), 176–188. https://doi.org/10.1007/s40817-020-00089-0

Mcauley, T., Chen, S., Goos, L., Schachar, R., & Crosbie, J. (2010). Is the behavior rating inventory of executive function more strongly associated with measures of impairment or executive function? *Journal of the International Neuropsychological Society*, *16*(3), 495–505. https://doi.org/10.1017/S1355617710000093

McCoy, D. C. (2019). Measuring young children's executive function and self-regulation in classrooms and other real-world settings. *Clinical Child and Family Psychology Review*, *22*(1), 63–74. https://doi.org/10.1007/s10567-019-00285-1

McDonald R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of PersonalityAssessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

Miranda, A., Colomer, C., Mercader, J., Fernández, M. I., & Presentación, M. J. (2015). Performance-based tests versus behavioral ratings in the assessment of executive functioning in preschoolers: Associations with ADHD symptoms and reading achievement. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00545

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Moens, M. A., Weeland, J., Van der Giessen, D., Chhangur, R. R., & Overbeek, G. (2018). In the eye of the beholder? Parent-observer discrepancies in parenting and child disruptive behavior assessments. *Journal of Abnormal Child Psychology*, *46*(6), 1147–1159. https://doi.org/10.1007/s10802-017-0381-7

Montgomery, D. E., & Koeltzow, T. E. (2010). A review of the day–night task: The Stroop paradigm and interference control in young children. *Developmental Review*, *30*(3), 308–330. https://doi.org/10.1016/j.dr.2010.07.001

Müller, U., & Kerns, K. (2015). The development of executive Function. In L. S. Liben & U. Müller (Eds.), *Handbook of child psychology and developmental science: Vol. 2. Cognitive processes* (7th ed., pp. 571-624). John Wiley & Sons. https://doi.org/10.1002/9781118963418.childpsy214

Nelson, T. D., Kidwell, K. M., Nelson, J. M., Tomaso, C. C., Hankey, M., & Espy, K. A. (2018).

Preschool executive control and internalizing symptoms in elementary school. *Journal of Abnormal Child Psychology*, *46*(7), 1509–1520. https://doi.org/10.1007/s10802-017-0395-1

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin*, *126*(2), 220–246. https://doi.org/10.1037/0033-2909.126.2.220

Nigg, J. T. (2017). Annual Research Review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *58*(4), 361–383. https://doi.org/10.1111/jcpp.12675

Nilsen, E. S., Huyder, V., McAuley, T., & Liebermann, D. (2017). Ratings of Everyday Executive Functioning (REEF): A parent-report measure of preschoolers' executive functioning skills. *Psychological Assessment*, *29*(1), 50–64. https://doi.org/10.1037/pas0000308

O'Meagher, S., Norris, K., Kemp, N., & Anderson, P. (2019). Examining the relationship between performance-based and questionnaire assessments of executive function in young preterm children: Implications for clinical practice. *Child Neuropsychology*, *25*(7), 899–913. https://doi.org/10.1080/09297049.2018.1531981

Pellicano, E., Kenny, L., Brede, J., Klaric, E., Lichwa, H., & McMillin, R. (2017). Executive function predicts school readiness in autistic and typical preschool children. *Cognitive Development*, *43*, 1–13. https://doi.org/10.1016/j.cogdev.2017.02.003

Petrides, M., & Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, *20*(3), 249–262.

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*, *87*(1), 102–112.

Rai, J. K., Abecassis, M., Casey, J. E., Flaro, L., Erdodi, L. A., & Roth, R. M. (2017). Parent rating of executive function in fetal alcohol spectrum disorder: A review of the literature and new data on Aboriginal Canadian children. *Child Neuropsychology*, *23*(6), 713–732. https://doi.org/10.1080/09297049.2016.1191628

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173–184. https://doi.org/10.1177/01466216970212006

Roth, R. M., Isquith, P. K., & Gioia, G. A. (2014). Assessment of executive functioning using the Behavior Rating Inventory of Executive Function (BRIEF). In S. Goldstein & J. A. Naglieri (Eds.), *Handbook of executive functioning* (pp. 301–331). Springer. https://doi.org/10.1007/978-1-4614-8106-5_18

Schmitt, S. A., Pratt, M. E., & McClelland, M. M. (2014). Examining the validity of behavioral self-regulation tools in predicting preschoolers' academic achievement. *Early Education and Development*, *25*(5), 641–660. https://doi.org/10.1080/10409289.2014.850397

Schneider, H., Ryan, M., & Mahone, E. M. (2020). Parent versus teacher ratings on the BRIEF-preschool version in children with and without ADHD. *Child Neuropsychology*, *26*(1), 113–128. https://doi.org/10.1080/09297049.2019.1617262

Schoemaker, K., Mulder, H., Deković, M., & Matthys, W. (2013). Executive functions in preschool children with externalizing behavior problems: A meta-analysis. *Journal of Abnormal Child Psychology*, *41*(3), 457–471. https://doi.org/10.1007/s10802-012-9684-x

Silver, C. H. (2014). Sources of data about children's executive functioning: Review and

commentary. *Child Neuropsychology*, *20*(1), 1–13. https://doi.org/10.1080/09297049.2012.727793

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K.E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.

Statistics Canada (2019). 2016 Census of population. Statistics Canada. https://www12.statcan.gc.ca/census-recensement/stats/statgeo2016.cfm?LANG=E&GEOCODE=01

Tamm, L., & Peugh, J. (2019). Concordance of teacher-rated and performance-based measures of executive functioning in preschoolers. *Child Neuropsychology*, *25*(3), 410–424. https://doi.org/10.1080/09297049.2018.1484085

Tan, A., Delgaty, L., Steward, K., & Bunner, M. (2018). Performance-based measures and behavioral ratings of executive function in diagnosing attention-deficit/hyperactivity disorder in children. *ADHD Attention Deficit and Hyperactivity Disorders*, *10*(4), 309–316. https://doi.org/10.1007/s12402-018-0256-y

Ten Eycke, K. D., & Dewey, D. (2016). Parent-report and performance-based measures of executive function assess different constructs. *Child Neuropsychology*, *22*(8), 889–906. https://doi.org/10.1080/09297049.2015.1065961

Thorell, L. B., Eninger, L., Brocki, K. C., & Bohlin, G. (2010). Childhood Executive Function Inventory (CHEXI): A promising measure for identifying young children with ADHD? *Journal of Clinical and Experimental Neuropsychology*, *32*(1), 38–43.

https://doi.org/10.1080/13803390902806527

Thorell, L. B., & Nyberg, L. (2008). The Childhood Executive Functioning Inventory (CHEXI): A new rating instrument for parents and teachers. *Developmental Neuropsychology, 33*(4), 536–552. https://doi.org/10.1080/87565640802101516

Thorndike, E. l. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner Review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry*, *54*(2), 131–143. https://doi.org/10.1111/jcpp.12001

Vries, M. de, Ruiter, M. A. de, Oostrom, K. J., Meeteren, A. Y. N. S.-V., Maurice-Stam, H., Oosterlaan, J., & Grootenhuis, M. A. (2018). The association between the behavior rating inventory of executive functioning and cognitive testing in children diagnosed with a brain tumor. *Child Neuropsychology*, *24*(6), 844–858. https://doi.org/10.1080/09297049.2017.1350262

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I Latent structure. *Developmental Psychology*, *44*(2), 575–587. https://doi.org/10.1037/0012-1649.44.2.575

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A. C., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, *108*(3), 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of Attention-Deficit/Hyperactivity Disorder: A meta-analytic review. *Biological Psychiatry*, *57*(11), 1336–1346.

https://doi.org/10.1016/j.biopsych.2005.02.006

Willoughby, M. T., Magnus, B., Vernon-Feagans, L., & Blair, C. B. (2017). Developmental delays in executive function from 3 to 5 years of age predict kindergarten academic readiness. *Journal of Learning Disabilities*, *50*(4), 359–372. https://doi.org/10.1177/0022219415619754

Willoughby, M. T., Wirth, R. J., & Blair, C. B. (2012). Executive function in early childhood: Longitudinal measurement invariance and developmental change. *Psychological Assessment*, *24*(2), 418–431. https://doi.org/10.1037/a0025779

Wolraich, M. L., Lambert, E. W., Bickman, L., Simmons, T., Doffing, M. A., & Worley, K. A. (2004). Assessing the impact of parent and teacher agreement on diagnosing Attention-Deficit Hyperactivity Disorder: *Journal of Developmental & Behavioral Pediatrics*, *25*(1), 41–47. https://doi.org/10.1097/00004703-200402000-00007

Young, A. R., Gurm, M. K., & O'Donnell, K. A. (2017). Assessing executive functions in young children. In M. J. Hosky, G. Iarocci, & A. R. Young (Eds.), *Executive functions in children's everyday lives: A handbook for professionals in applied psychology* (pp. 21–37). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199980864.003.0003

Zelazo, P. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*, 297–301. https://doi.org/10.1038/nprot.2006.46

Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives*, *6*(4), 354–360. https://doi.org/10.1111/j.1750-8606.2012.00246.x

Zelazo, P. D., & Müller, U. (2002). Executive function in typical and atypical development. In U.

Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 445–469). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470996652.ch20