An Exploratory Study Regarding the Ease-of-use, Comprehensibility, and Usefulness of the Empirical Standards Checklists

by

Cassandra Cupryk
B.Sc., University of Toronto, 2019

A Master's Project Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Supervisory Committee

Dr. Margaret-Anne Storey, Supervisor
(Department of Computer Science)

Dr. Neil Ernst, Departmental Member
(Department of Computer Science)

ABSTRACT

Context/Background: Novice researchers have stated that being provided with guidelines for reviewing empirical research papers would be helpful. The Empirical Standards Checklist Generator is a tool that can generate a variety of Empirical Standards Checklists. An Empirical Standards Checklist contains the core criteria that can be used to review Software Engineering empirical papers. Moreover, my exploratory study aims to determine whether novice researchers could benefit from using the Empirical Standards Checklists to help them review Software Engineering empirical papers.

Objective: To investigate whether novice researchers perceive the Empirical Standards Checklists as easy to understand, easy to use, and useful for reviewing empirical papers.

Methods: Seven participants completed a survey to evaluate the Empirical Standards Checklists and then participated in a group discussion. During the survey, the participants used the appropriate Empirical Standard Checklists to review a qualitative survey paper and a repository mining paper. They then highlighted the items from the Empirical Standards Checklists that were difficult to comprehend. The participants also answered survey questions that exposed their perceptions of the Empirical Standards Checklists' comprehensibility, ease-of-use, and usefulness.

Results: The majority of the participants had positive perceptions of the Empirical Standards Checklists' comprehensibility, ease-of-use, and usefulness.

Conclusion: This exploratory study demonstrates that the Empirical Standards Checklist is a promising new tool for reviewing Software Engineering empirical research papers.

Contents

Supervisory Committee **ii**

Abstract **iii**

Table of Contents **iv**

List of Tables **vii**

List of Figures **viii**

Acknowledgements **x**

Dedication **xi**

1 Introduction **1**

1.1 Motivation 1

1.2 Research Questions 3

1.3 Structure of the Report 4

1.4 Report's Acronyms and Corresponding Terms 4

2 Related Works **5**

2.1 Surveys 6

2.2 Case Study 7

2.3 Experiments 7

2.4 The Empirical Standards Checklist Generator (ESCG) 9

2.4.1 History of the Empirical Standards Checklists of the ESCG 10

2.4.2 How to use the ESCG for reviewing papers 10

3 Methodology **19**

3.1 Study Design (for Pilot, Session 1, and Session 2) 20

3.1.1 Survey Construction 20
3.1.2 Paper Selection 26
3.1.3 Participant Selection 26
3.1.4 Ethics 27
3.1.5 Recruitment for Pilot and Feedback from Pilot 27
3.1.6 Recruitment for Sessions of My Study 28
3.2 Process for Pilot and Each Study Session 29
3.2.1 Part 1 of Pilot and Each Study Session: Survey 30
3.2.2 Part 2 of Pilot and Each Study Session: Group Discussion 32
3.3 After the Pilot and Sessions of My Study 32
3.4 Theme Code Analysis 33

4 Results **34**
4.1 Participants' Demographic Data 34
4.2 Accepting or Rejecting Papers 38
4.3 RQ1: What are the Ph.D. and Master's students' perceptions concerning the **comprehensibility** of the Empirical Standards Checklist? 40
4.3.1 Participants' **General** Perceptions on the Comprehensibility of the Empirical Standards Checklists 40
4.3.2 Participants' Perceptions on the Comprehensibility of the **Qualitative Survey** Empirical Standards Checklist 40
4.3.3 Participants' Perceptions on the Comprehensibility of the **Repository Mining** Empirical Standards Checklist 42
4.4 RQ2: What are the Ph.D. and Master's students' perceptions concerning the **ease-of-use** of the Empirical Standards Checklist? 44
4.5 RQ3: What are the Ph.D. and Master's students' perceptions concerning the **usefulness** of the Empirical Standards Checklist? 46
4.5.1 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to review research papers more quickly? 47
4.5.2 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to review research papers more easily? 47

4.5.3 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to increase their understanding of research papers? . . . 48
4.5.4 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to improve how they review research papers? . . . 48

5 Discussion **50**
5.1 An Implication for Ph.D. and Master's students . . . 50
5.2 Recommendations for the Empirical Standards Checklist . . . 50
5.2.1 Recommendations for the UI Designers of the Empirical Standards Checklists . . . 51
5.2.2 Recommendations for the Research Community of Empirical Standards Checklist . . . 51
5.3 Validity threats and limitations . . . 52
5.3.1 Internal Validity . . . 52
5.3.2 Construct validity . . . 52
5.3.3 Reliability . . . 52

6 Conclusion **53**

A Additional Information **54**
A.1 Final Theme Codes . . . 55
A.2 Survey for Pilot Study . . . 56
A.3 Survey for Studies Other Than Pilot Study . . . 69
A.4 Microsoft Form for Recruitment of Participants . . . 81
A.5 Research Paper for Study - Qualitative Survey Paper . . . 85
A.6 Research Paper for Study - Repository Mining Paper . . . 97
A.7 Ethics - Session 1 - Email to Participants . . . 109
A.8 Ethics - Session 1 - Consent Form to be Signed by Participants . . . 112
A.9 Ethics - Session 2 - Email to Participants Including Implied Consent Form 115
A.10 Ethics - HREB Application Approval - Session 1 . . . 118
A.11 Ethics - HREB Application Approval - Session 2 . . . 138
A.12 Ethics - TCPS2 Core Certificate for Cassandra Cupryk . . . 160

Bibliography **162**

List of Tables

Table 1.1 Acronyms for Terms in the Report 4

Table 4.1 Demographic Data - Participants' Information 35

Table A.1 Theme Codes for the Qualitative Data Analysis. 55

List of Figures

Figure 2.1 ESCG - Homepage 11
Figure 2.2 ESCG - List of Empirical Methodologies 13
Figure 2.3 ESCG - Essential Section of Qualitative Survey Checklist 15
Figure 2.4 ESCG - Desired Section of Qualitative Survey Checklist 16
Figure 2.5 ESCG - Extraordinary Section of Qualitative Survey Checklist . . . 17
Figure 2.6 ESCG - Example of reviewChecklist.txt file for Qualitative Survey Checklist 18

Figure 3.1 Timeline of Study: Pilot, Session 1 of Study, Session 2 of Study . 19
Figure 3.2 Molléri et al. - Perceptions of Comprehensibility Survey Questions 21
Figure 3.3 Section of Survey - Comprehensibility of the Empirical Standards Checklist for Reviewing Qualitative Survey Paper 21
Figure 3.4 Section of Survey – Ease-of-Use of Empirical Standards Checklist for Qualitative Survey Paper 22
Figure 3.5 Davis - Final Scale to Measure User's Perceptions of Usefulness and Ease-of-Use for Chart Master 23
Figure 3.6 Davis - Perceptions of Usefulness Survey Questions 24
Figure 3.7 Section of Survey - Usefulness of Empirical Standards Checklist for Reviewing Qualitative Survey Paper 25
Figure 3.8 Recruitment Post for Session 1 of My Study 29
Figure 3.9 Recruitment Post for Session 2 of My Study 30
Figure 3.10 Example of reviewChecklist.txt file 31

Figure 4.1 Demographic Data - Overview of the Empirical Standard Methodologies that are Familiar to the Participants 36
Figure 4.2 Demographic Data - Results of Demographic Survey Questions . . 37
Figure 4.3 Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the Repository Mining Paper 39

Figure 4.4 Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the Qualitative Survey Paper 39
Figure 4.5 Comprehensibility - Results for the Survey Question: "The Statements from the Empirical Standards Checklist for the Qualitative Survey Paper were Overall Easy to Understand" 41
Figure 4.6 Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the Qualitative Survey Paper 42
Figure 4.7 Comprehensibility - Results for the Survey Question: "The Statements from the Empirical Standards Checklist for the Repository Mining Paper were Overall Easy to Understand" 43
Figure 4.8 Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the RM Paper 44
Figure 4.9 Ease-of-Use - Results for the Survey Question: "Overall, I find the Empirical Standards Checklists Easy to Use" 45
Figure 4.10 Usefulness - Overview of Quantitative Usefulness Results of the Empirical Standards Checklist for the Qualitative Survey Paper 46
Figure 4.11 Usefulness - Overview of Quantitative Usefulness Results of the Empirical Standards Checklist for the Qualitative Survey Paper 47

ACKNOWLEDGEMENTS

I would like to thank:

my sister, Sabrina for gifting me some excellent life advice that helped me overcome a gigantic mental hurdle.

my sister's dog, Misty, for providing me with countless cuddles.

my parents, for providing me with delicious food and endless support.

Dr. Paul Ralph, for letting me be a part of the Empirical Standards Checklists Generator project.

The Members of C.H.I.S.E.L, for helping me in a billion different ways.

Dr. Margaret-Anne D. Storey for teaching me a lot about the importance of qualitative research, for allowing me to finish at my own pace, and for helping me gain my confidence back before I enter the industry.

Enrique Larios Vargas for taking a lot of time out of their busy schedule to aid me with designing my project, conducting my project, and writing up my project.

"Letting it get to you. You know what that's called? Being alive. Best thing there is. Being alive right now is all that counts."
The Eleventh Doctor from Doctor Who

DEDICATION

Thank you to everyone who has worked on and who is currently working on the Empirical Standards Checklists Generator.
I can't wait to see what becomes of this magnificent tool.

Chapter 1

Introduction

1.1 Motivation

In "Manuscript reviewing: What reviewers have to say", Emden et al. described a study where they interviewed fifteen reviewers to collect their perceptions on the strengths and weaknesses of the review process, their experiences with the review process, and which areas of the review process they believe need improvement. [2]. During these interviews, several reviewers stated they lacked preparation and direction in reviewing a manuscript. One respondent stated "I have not had any orientation to the reviewing process and feel that 'an interest in an area' is not enough to be making decisions re the merit about someone else's work." In their study, Emden et al. also found that reviewers were frustrated due to the various editors' expectations that differed among all the various journals. Reviewers were also confused by the lack of clarity in the guidelines outlined by each journal. Some reviewers stated that they were confused about whether they had to verify every reference and how detailed they had to be when they edited a manuscript. One novice reviewer stated that "Some general guidelines as to how to go about the review process would have been helpful[2]." Moreover, reviewers felt that the editors could better prepare reviewers by providing them with clearer guidelines [2].

In a related study, "Manuscript Peer Review: A Helpful Checklist For Students And Novice Referees", Seals et al. points to references highlighting that novice researchers do not have much experience reviewing research papers [17]. Lightfoot stated that through their personal experience, they found that both undergraduate and graduate students had a poor understanding of how to review papers due to having never been taught how to review papers [11]. Rangachari et al. state that graduate students are used to reading

textbooks and taking notes but are unfamiliar with performing critical analyses of papers. However, Seals et al. state that developing students' reviewing skills is imperative to prepare them for future editorial responsibilities [17].

To address the issues noted above, checklists could be employed to aid novice researchers, or more specifically Ph.D. and Master's students, in reviewing Software Engineering (SE) empirical papers. This solution is supported by the fact that two papers have presented that other students would benefit from having guidelines to review papers. In "Manuscript Peer Review: A Helpful Checklist For Students And Novice Referees", Seals et al. presented a checklist for graduate students and postdoctoral individuals to review research papers critically. Seals et al. built this checklist because they felt that students and novice researchers tend to focus on the methodological limitations of a research paper when reviewing papers rather than on other key criteria. The feedback that Seals et al. received was positive for the guidelines since the students stated that the guidelines were a "very helpful guide".

In "A Checklist To Help Students Analyze Published Articles in Basic Medical Sciences", Rangachari et al. developed a checklist of questions separated into Introduction, Methodology, Research, And Discussion (IMRAD) sections that undergraduate students used to review Pharmacology research papers [16]. As part of this work, students were expected to read a Pharmacology research paper and then write a report critically evaluating the paper. The participants then provided their insights on the checklist. For instance, one participant stated, "A large part of what we do is evaluate the quality of the papers we read. The checklist gave formal instructions important for someone who has never critically analyzed a paper and allowed it to be broken down into manageable parts. I've used the checklist when I needed to analyze papers for other courses and found it to be really helpful." Another participant added, "This was a very effective exercise, both in terms of analysis of a paper and as a help in writing your own papers." However, there were negative views of the checklist as well. One participant stated that "Critical analysis arises from an overall perception, and I believe that the formality of the process, complete with checklist, biases perception." Moreover, Rangachari was able to obtain both positive and negative commentaries about the checklist to demonstrate the checklist's value and demonstrate the checklist has room for improvement [16].

From these findings that showed that students felt that guidelines would help them review papers, in my research, I explore the use of a tool called the "Empirical Standards Checklist Generator (ESCG) [1]" designed by Dr. Paul Ralph, a Professor of Software

[1] https://acmsigsoft.github.io/EmpiricalStandards/tools/

Engineering at Dalhousie University in Halifax, Canada. The ESCG is implemented as a website that contains a repository of core electronic checklists called the "Empirical Standards Checklists (ESCs)" to help authors and reviewers write and review a variety of SE empirical papers. In addition, in "Empirical Standards for Software Engineering Research", Dr. Ralph provides details on the design process for the ESCG, how the ESCG can be adopted into the research community, the benefits and costs of the ESCG, and how to mitigate bias and inclusivity in the ESCG [15].

However, before the ESCG is introduced into the research community, an exploratory study may be beneficial to examine whether the ESCG is desired by the SE research community. In "Development of medical checklists for improved quality of patient care", Hales et al. state that running a pilot for a checklist is necessary to determine whether the checklist is practical, effective, and needed[6]. In my research, I conduct an exploratory study with ECSG. My exploratory study is split into two main parts. The first part consists of a survey, where the participants use two ESCs to review two empirical papers and answer several survey questions to provide their perceptions of the comprehensibility, ease-of-use, and usefulness of the two ESCs. The second part of the study involves a group discussion with the surveyed participants to allow them to provide further opinions on the checklists that they could not share through the survey.

1.2 Research Questions

In order to determine Ph.D. and Master's students' perceptions of the comprehensibility, ease-of-use, and usefulness of the ESCs, my work focuses on answering these three research questions:

1. (RQ1) What are Ph.D. and Master's students' perceptions concerning the **comprehensibility** of the Empirical Standards Checklists?
2. (RQ2) What are Ph.D. and Master's students' perceptions concerning the **ease-of-use** of the Empirical Standards Checklists?
3. (RQ3) What are Ph.D. and Master's students' perceptions concerning the **usefulness** of the Empirical Standards Checklists?

1.3 Structure of the Report

This report is organized as follows:

In Chapter 1, I introduce the research questions and the motivation for undertaking this exploratory study.

In Chapter 2, I elaborate on several related works to address the challenges students face when doing research.

In Chapter 3, I outline the study's methodology in enough detail in case other researchers may want to replicate my study.

In Chapter 4, I present the Master's and Ph.D. students' perceptions on the comprehensibility, ease-of-use, and usefulness of the Empirical Standards Checklist.

In Chapter 5, I discuss the implications and limitations of the results.

In Chapter 6, I address the potential future work of this study as well summarize the outcomes of this study.

1.4 Report's Acronyms and Corresponding Terms

Acronym	Term
ESCG	Empirical Standards Checklist Generator
ESC	Empirical Standards Checklist
QS	Qualitative Survey
RM	Repository Mining
SE	Software Engineering
CS	Computer Science

Table 1.1: Acronyms for Terms in the Report

Chapter 2

Related Works

Before I did the study, I wanted to make sure there wasn't something similar to the ESCG that already existed. Thus, I looked for some software, tool, framework, or checklist that could help one review software engineering empirical papers, specifically a repository mining (RM) paper and a qualitative survey (QS) paper since these two empirical papers would be reviewed during my study. I started by looking for papers that discussed software or tools that could help researchers conduct reviews of empirical papers in the following databases: "ACM digital library", "ARXIV", "Elsevier ScienceDirect", "Google Scholar", and "IEEE Xplore Digital Library". For some papers, I performed snowballing to see if I could discover more papers that included tools or software to review software engineering empirical papers. In addition, for some searches, I had to omit the terms "systematic" and "literature" in order to prevent papers containing systematic literature reviews from appearing in all my searches. I used the University of Victoria (UVic) library in order to access the papers.

While there were tools that helped researchers **conduct** empirical methodologies, specifically systematic literature reviews [3] [12], I was unable to find any software or tools that would help reviewers (or students) conduct reviews of software engineering empirical papers.

Through my literature search, I found "Text & Vision-Fused Framework for Academic Paper Review", where Yang et al. proposed the first framework (they claim) would help filter out the low-quality research papers to save the review committee time when reviewing papers. The framework is based on a deep-learning model that considers the quality of the paper's vocabulary and images. This framework can be used to estimate the probability of a research paper being accepted [21]. The results of Yang et al.'s paper demonstrate that their framework was effective, as it was 98.7% accurate when

rejecting lower-quality papers with a small mis-reject percentage of 4%. Even though Yang et al. have created an effective new framework for automating the acceptance and rejection of papers, the framework is still not 100% accurate, which could pose issues when deciding whether to accept or reject empirical papers. In addition, Yang et al. also state that they only tested their framework on a small dataset of conference papers from only one conference (OpenReview) and thus, their results cannot be generalized to determine whether to accept or reject papers in other conferences. Moreover, Yang et al's framework still needs to be improved before it can be integrated into the Software Engineering research community.

There are several other papers that provide guidelines and checklists that claim to help in reviewing empirical papers with the following four methods: surveys [14], case studies [7] [19], and experiments [8] [9] [19] [20].

2.1 Surveys

In "An empirically evaluated checklist for surveys in software engineering," published in 2020, Molléri et al. propose a checklist that can be beneficial for researchers to conduct and report survey research papers and for reviewers who would like to review survey research papers. The checklist was created by merging guidelines from fourteen different papers (referenced in this paper). Selected experienced researchers then evaluated the checklist to determine the checklist's comprehensibility and limitations. The selected researchers were tasked with reviewing a set of survey papers with the checklist and to provide their ratings of these papers. The researchers' results were compared with the authors' evaluations of the selected survey papers. Nineteen out of the thirty-eight checklist items were improved due to the researchers' feedback in the study. The final version of the checklist following this phase had ten sections as follows: (1) Research Objectives, (2) Study Plan, (3) Identify population, (4) Sampling Plan, (5) Instrument Design, (6) Instrument Validation, (7) Participant Recruitment, (8) Response Management, (9) Data Analysis and (10) Reporting. Molléri et al.'s checklist may be useful for reviewing a qualitative survey paper. However, Molléri did not construct a checklist to review a repository mining paper. Moreover, the checklist can only help reviewers review papers with one type of method.

2.2 Case Study

In "Checklists for Software Engineering Case Study Research," published in 2007, Höst constructed a checklist by combining several checklists. Höst then asked Ph.D. students to validate the checklist within a course. Höst tasked the Ph.D. students to highlight the checklist items that were difficult for the students to understand and the useful checklist items. The Ph.D. students informed Höst that the checklist was too extensive. As a result, Höst determined how to group the checklist items and reduce the overall checklist items in the checklist. After further validating the checklist with the Ph.D. students, Höst et al. obtained two checklists. One checklist was designed for researchers and was divided into four sections: (1) case study design, (2) preparation for data collection, (3) analysis of collected data, and (4) reporting. The other checklist was aimed at reviewers, where each checklist item corresponded to checklist items in the researcher's checklist. Even though Höst's study produced a validated checklist for reviewers, this checklist can only be used to review one empirical method: case studies.

2.3 Experiments

In "Preliminary Guidelines for Empirical Research in Software Engineering," published in 2002, Kitchenham et al. created a set of quality guidelines for planning experiment studies as well as reading, writing, and reviewing experiment papers [8]. Kitchenham et al. also stated that the guidelines could help a meta-analyst combine the knowledge from a variety of studies and help journal editorial boards decide on whether to accept or reject a paper. Kitchenham et al. crafted these guidelines by adapting them from various medical guidelines. The authors specifically selected medical guidelines, since medical statisticians have been diligent in identifying poor standards of statistical analysis in medical journals. Thus, medical guidelines include checklist items that examine the statistical analysis in medical journals. Kitchenham et al. also found that many SE empirical papers have had issues correctly reporting statistical data. Moreover, Kitchenham claims if researchers used these empirical guidelines adapted from the medical guidelines, the quality of papers could increase. Kitchenham et al.'s finalized guidelines were split into six main sections:(1)"Experimental Context", (2) "Experimental Design", (3) "Conducting the experiment and data collection" (4) "Analysis", (5) "Presentation of Results", and (6)"Interpretation of Results". Even though Kitchenham et al. proposed detailed guidelines for reviewing experiment papers, these guidelines are useful for reviewing one

empirical method (experiments).

In "Can We Evaluate the Quality of Software Engineering Experiments?", published in 2010, Kitchenham et al. aimed to achieve two primary goals. One goal was to verify the usability of their constructed quality review checklist for experiment papers. Another goal was to determine how many reviewers needed to be included in the review process to provide a reliable evaluation of an experiment paper. The quality checklist was a modified version of various checklists with three categories: (1) questions on aims, (2) questions on design, data collection, and data analysis, and (3) questions on study outcome. Kitchenham et al. organized several rounds of study where experienced researchers iteratively validated the quality of the checklist. After the quality of the checklist was validated, the authors then used the quality checklist to determine the number of reviewers needed to provide reliable evaluations of experiment papers. From their findings, four reviewers, collaborating in pairs, resulted in a higher reliability of determining the quality of a paper compared to eight reviewers who did not discuss their paper evaluations. Even though Kitchenham et al. produced a new quality checklist that underwent several rounds of evaluation, the checklist can only be used for reviewing one type of empirical paper: experiments.

In the book published in 2012, "Experimentation in Software Engineering," Wohlin et al. provide extensive detail on how to report and review experiment papers for practitioners, students, teachers, and researchers. Wohlin et al. also present some information on some other empirical methodologies, such as case studies, systematic literature reviews, and surveys, and how they relate to experiments. Wohlin et al. state that their book fills a gap in the research since it focuses on the experiment process instead of just focusing on one specific portion of experiments, for instance, discussing the statistical methods in experiments. Wohlin et al. present the steps in the experiment process. (1) "Scoping" outlines the experiment's goals according to a specific framework. (2) "Planning" involves figuring out how the experiment will run by following seven steps: context selection, hypothesis formulation, variable selection, selection of subjects, experiment design type, instrumentation, and validity evaluation. (3) "Operation" is when the subjects are selected, the subjects perform the pre-defined tasks, and the data is validated. (4) "Analysis and Interpretation" describes how conclusions are obtained from examining experimental quantitative data using descriptive statistics, reducing the data set, and hypothesis testing. (5) "Presentation and Package" focuses on how to write a report for a journal or conference. Moreover, even though Wohlin et al. have provided extensive information in their book, they only cover experiments and how case studies, systematic

literature reviews, and surveys are related to experiments.

In "Towards a Unified Checklist for Empirical Research in Software Engineering: First proposal," Wieringa [19] published in 2012, Wieringa aimed to create a unified checklist that would integrate the similarities and differences that occur among experimental research and case study research. To create this checklist, Wieringa adapted the "general engineering cycle" to create their empirical research cycle, which was divided into five sections: research problem investigation, research design, research design validation, research execution, and results evaluation. Wieringa created their unified checklist using their empirical research cycle as a foundation and included checklist items from various experiment and case study checklists. Wieringa et al. used their unified checklist within their research group to teach empirical research methods to Master's and Ph.D. students. In the future, Wieringa claimed they would like to obtain researchers' opinions on the usability and usefulness of the checklist. One of the limitations of their checklist is their guidelines are primarily for reviewing experimental and case study research papers and cannot be used to review other empirical papers using other methods.

2.4 The Empirical Standards Checklist Generator (ESCG)

While there is one framework that shows promise in automating the process of reviewing papers, and there are several checklists that can educate individuals on how to review various software engineering empirical methods, there is not currently a software or a tool that has guidelines that can help individuals review empirical papers. The Empirical Standards Checklist Generator (ESCG) designed by Dr. Paul Ralph, a Professor of Software Engineering at Dalhousie University, fills this gap. Not only is the ESCG claimed to be a tool that can help other individuals review empirical research papers, but the ESCG contains checklists, which are called "the Empirical Standards Checklists", for reviewing **eighteen** different types of empirical papers (including checklists for reviewing a Repository Mining (RM) paper and Qualitative Survey (QS) paper) in one unified location. In addition, the extensive knowledge for reviewing empirical methods presented in these eighteen checklists would not be available if it weren't for the efforts of experienced researchers in various fields.

2.4.1 History of the Empirical Standards Checklists of the ESCG

In 2019, the ESCG was first launched. Dr. Ralph was able to recruit 50 volunteers to begin working on the Empirical Standards Checklists of the ESCG. Volunteers reviewed plans to create an Empirical Standards Checklist, discussed which Software Engineering methods should have an Empirical Standards Checklist, and selected the content for the sections of an Empirical Standards Checklist [15]. Volunteers also suggested other volunteers who might be interested in working on the Empirical Standards Checklists, and nominated the volunteers who would write the content for an Empirical Standards Checklist [15].

All the checklists were then circulated within the Software Engineering (SE) research community to obtain feedback and improve the content for the Empirical Standard Checklists [15]. Once the first eight Empirical Standard Checklists were posted on a GitHub repository, even more feedback was collected from the SE research community. As time went on, more experienced volunteer researchers were recruited to generate more Empirical Standard Checklists. Today, there are eighteen Empirical Standards Checklists that can be used to evaluate Software Engineering empirical papers presented in Figure 2.2. In addition, the Empirical Standard Checklists are currently available as Markdown files to allow volunteer researchers to easily make edits to the checklists and maintain version control [15].

2.4.2 How to use the ESCG for reviewing papers

To demonstrate how to use the ESCG, I present an example of how to generate a checklist to review a Qualitative Survey (QS) paper, as my research study involves a conference paper with a qualitative survey method. When visiting the homepage of the ESCG, presented in Figure 2.1, a user is faced with two main choices. One choice is a pre-submission checklist which is primarily for authors and can generate a checklist that will allow authors to check their papers before submission for review. The other choice is for reviewers and allows reviewers to generate a checklist to aid them when reviewing two types of papers: conference papers or journal papers (as my study involves a conference paper, I select that option in this illustrative example).

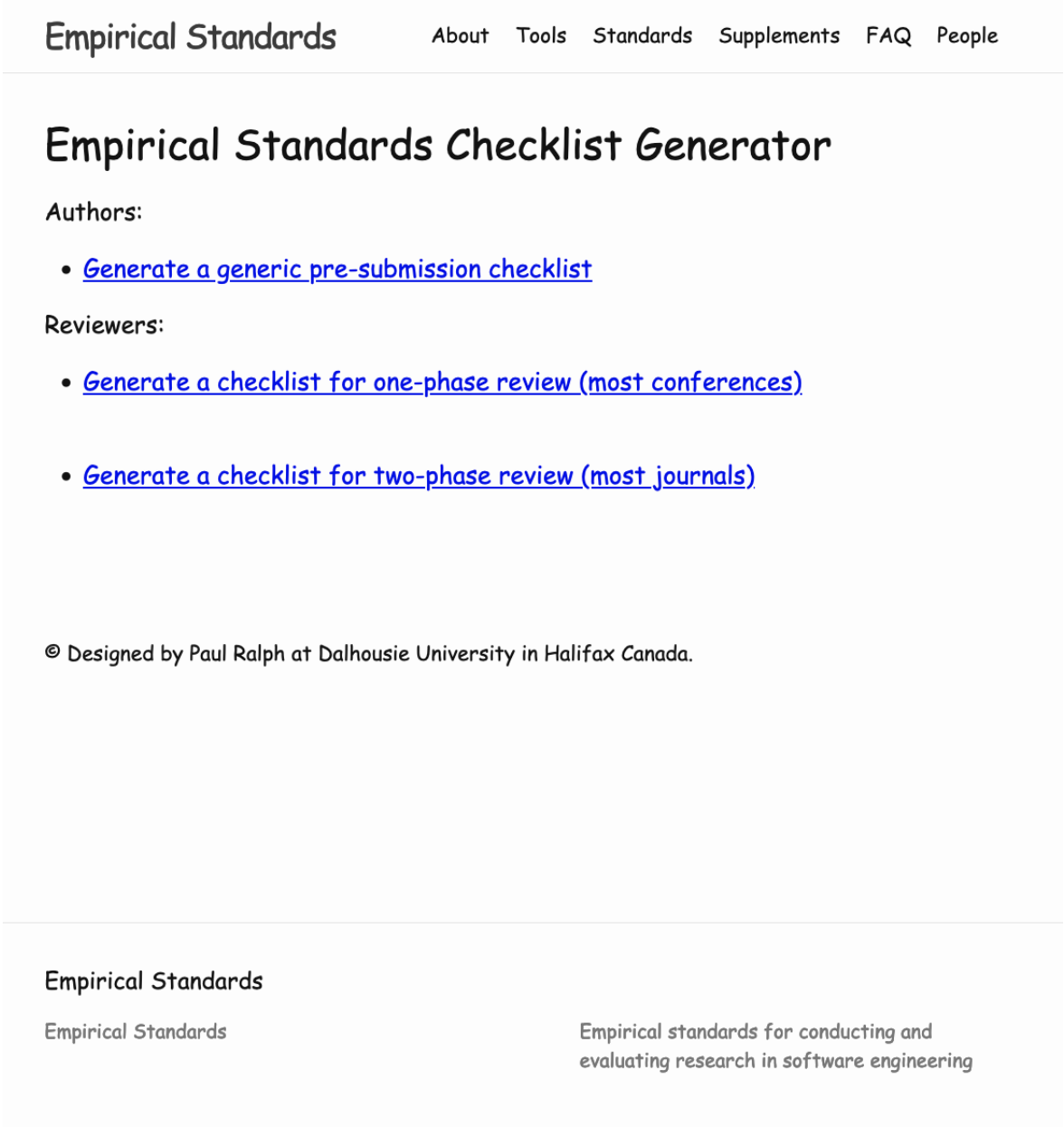


Figure 2.1: ESCG - Homepage

The next web page presented by the checklist generator tool consists of eighteen types of empirical methods that a user can choose from, shown in Figure 2.2. These types are listed below. Once one (or more) of these methods are selected, then a checklist tailored to the methods selected is generated.

1. Action research
2. Benchmarking

3. Case Study
4. Case Survey
5. Data Science
6. Engineering Research
7. Experiment (with human participants)
8. Grounded theory
9. Meta-science
10. Multi-methodology or mixed methods
11. Optimization study (including search-based software engineering)
12. Qualitative survey (i.e., interviews)
13. Quantitative longitudinal study
14. Quantitative simulation
15. Questionnaire survey
16. Repository Mining
17. Systematic Literature Reviews
18. An empirical method that is not currently listed (or General Standards)

Figure 2.2: ESCG - List of Empirical Methodologies

The next page in the generator tool consists of a checklist of core criteria (presented in the form of checklist items) for the previously selected empirical method(s) and is divided into three sections ("Essential", "Desirable", "Extraordinary"). The checklist items located in the "Essential" section are the checklist items that need to be found in the paper when the paper is being reviewed. A checklist item can be either marked "Yes" or "No". If a checklist item has been found in the paper, then the checklist item is marked "Yes". If the checklist item is not found in the paper, then the checklist item is marked "No". The "Essential" section is what primarily determines whether to accept or reject the paper. Moreover, once all the checklist items in the "Essential" section have been checked either "Yes" or "No", the checklist will output a message in red that either says "ACCEPT", or "REJECT". The checklist items located in the "Desirable" and "Extraordinary" sections are checklist items that would be ideal for the paper to have but are not necessary.

Continuing with the example from my research study, the checklist to review a QS paper is separated into three sections, as shown in Figure 2.3, 2.4, 2.5. In addition, Figure 2.3 shows that all the "Essential" checklist items have been marked "Yes". Thus, the message at the end of the checklist states "ACCEPT", meaning that the paper will be accepted.

If the user would like to keep a record of the checklist, then they would click the "Download" button at the bottom of the page (demonstrated in Figure 2.5) and a file titled "reviewChecklist.txt" would be downloaded to their device. An example of the downloaded "reviewChecklist.txt" for the Qualitative Survey Checklist is shown in Figure 2.6.

As demonstrated by the example from my research study, the ESCG can successfully generate ESCs that can be used to review SE empirical papers, but to date, little evaluation of the ESCG has been done. Moreover, my research focuses on a preliminary study of the ESCG, which I will describe in the following chapters.

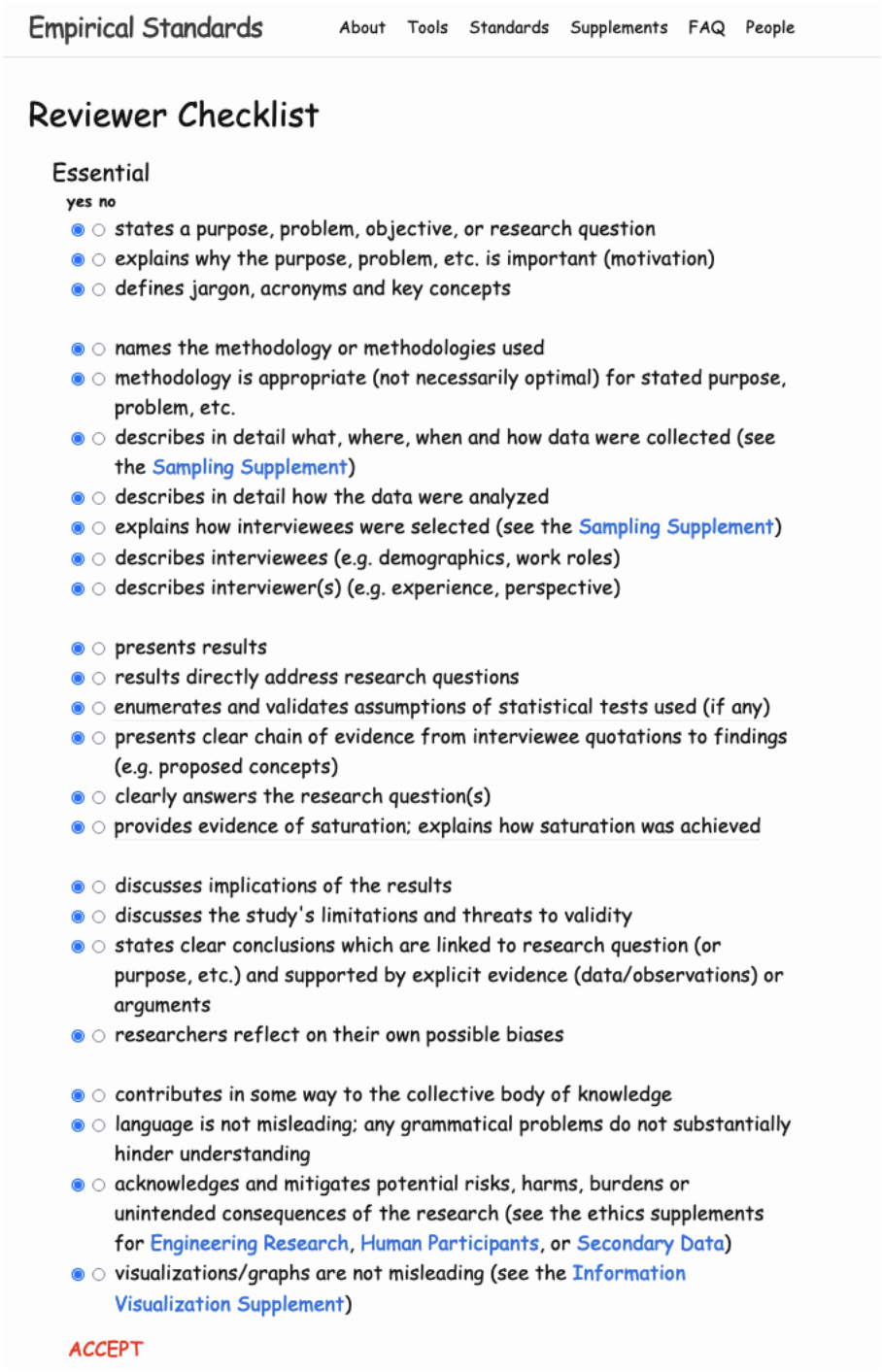
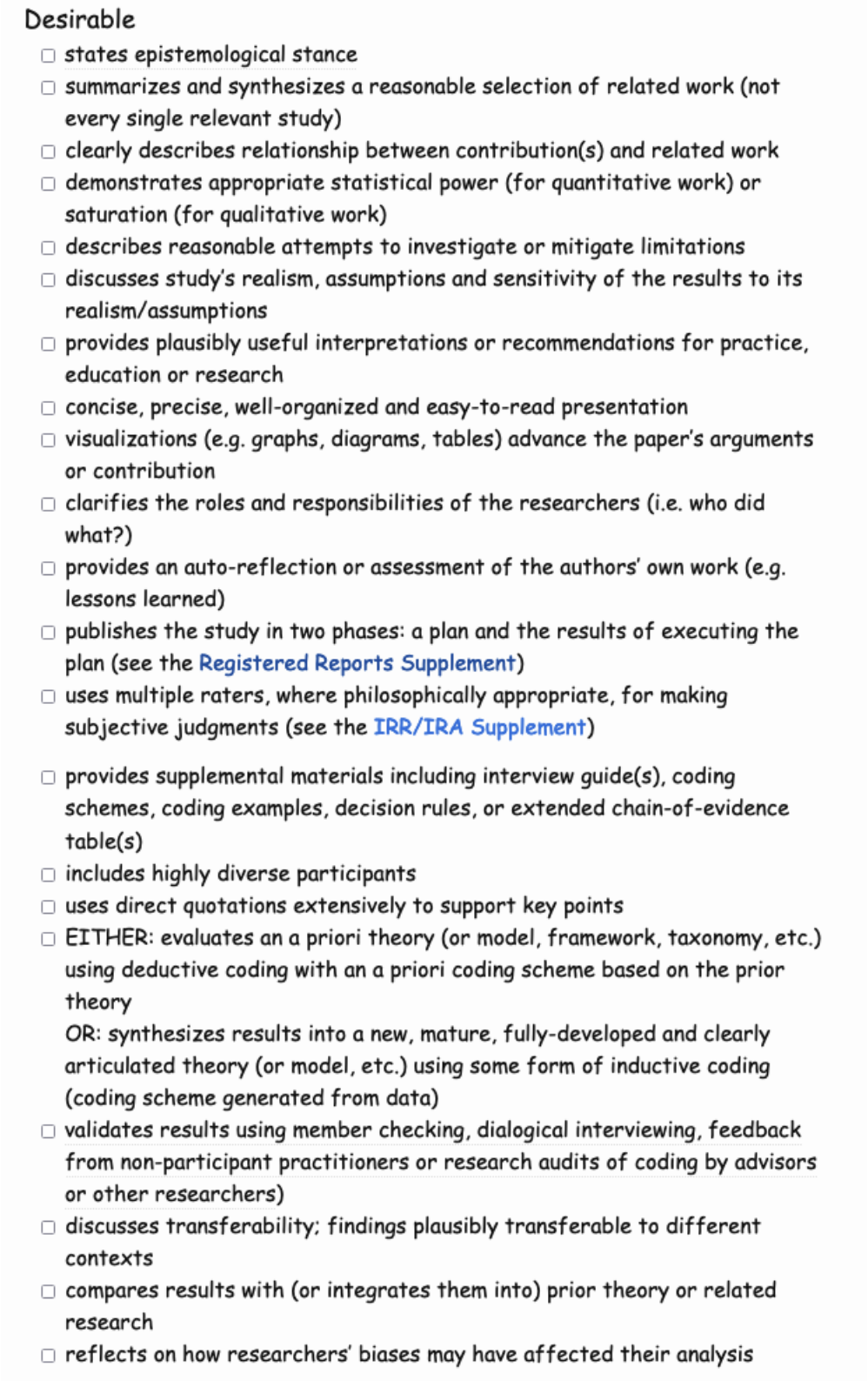


Figure 2.3: ESCG - Essential Section of Qualitative Survey Checklist



Desirable

- ☐ states epistemological stance
- ☐ summarizes and synthesizes a reasonable selection of related work (not every single relevant study)
- ☐ clearly describes relationship between contribution(s) and related work
- ☐ demonstrates appropriate statistical power (for quantitative work) or saturation (for qualitative work)
- ☐ describes reasonable attempts to investigate or mitigate limitations
- ☐ discusses study's realism, assumptions and sensitivity of the results to its realism/assumptions
- ☐ provides plausibly useful interpretations or recommendations for practice, education or research
- ☐ concise, precise, well-organized and easy-to-read presentation
- ☐ visualizations (e.g. graphs, diagrams, tables) advance the paper's arguments or contribution
- ☐ clarifies the roles and responsibilities of the researchers (i.e. who did what?)
- ☐ provides an auto-reflection or assessment of the authors' own work (e.g. lessons learned)
- ☐ publishes the study in two phases: a plan and the results of executing the plan (see the Registered Reports Supplement)
- ☐ uses multiple raters, where philosophically appropriate, for making subjective judgments (see the IRR/IRA Supplement)
- ☐ provides supplemental materials including interview guide(s), coding schemes, coding examples, decision rules, or extended chain-of-evidence table(s)
- ☐ includes highly diverse participants
- ☐ uses direct quotations extensively to support key points
- ☐ EITHER: evaluates an a priori theory (or model, framework, taxonomy, etc.) using deductive coding with an a priori coding scheme based on the prior theory
  OR: synthesizes results into a new, mature, fully-developed and clearly articulated theory (or model, etc.) using some form of inductive coding (coding scheme generated from data)
- ☐ validates results using member checking, dialogical interviewing, feedback from non-participant practitioners or research audits of coding by advisors or other researchers)
- ☐ discusses transferability; findings plausibly transferable to different contexts
- ☐ compares results with (or integrates them into) prior theory or related research
- ☐ reflects on how researchers' biases may have affected their analysis

Figure 2.4: ESCG - Desired Section of Qualitative Survey Checklist

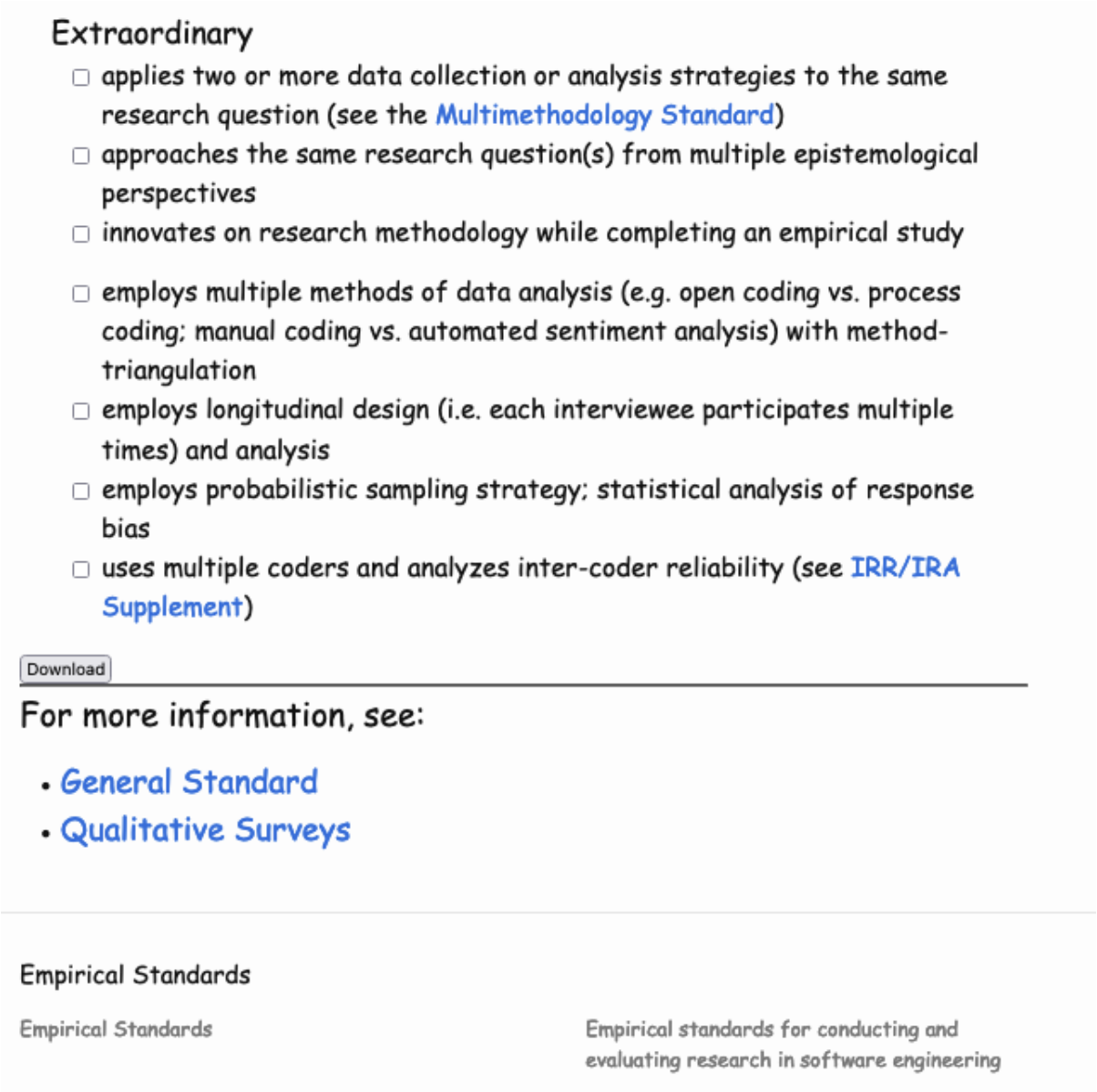


Figure 2.5: ESCG - Extraordinary Section of Qualitative Survey Checklist

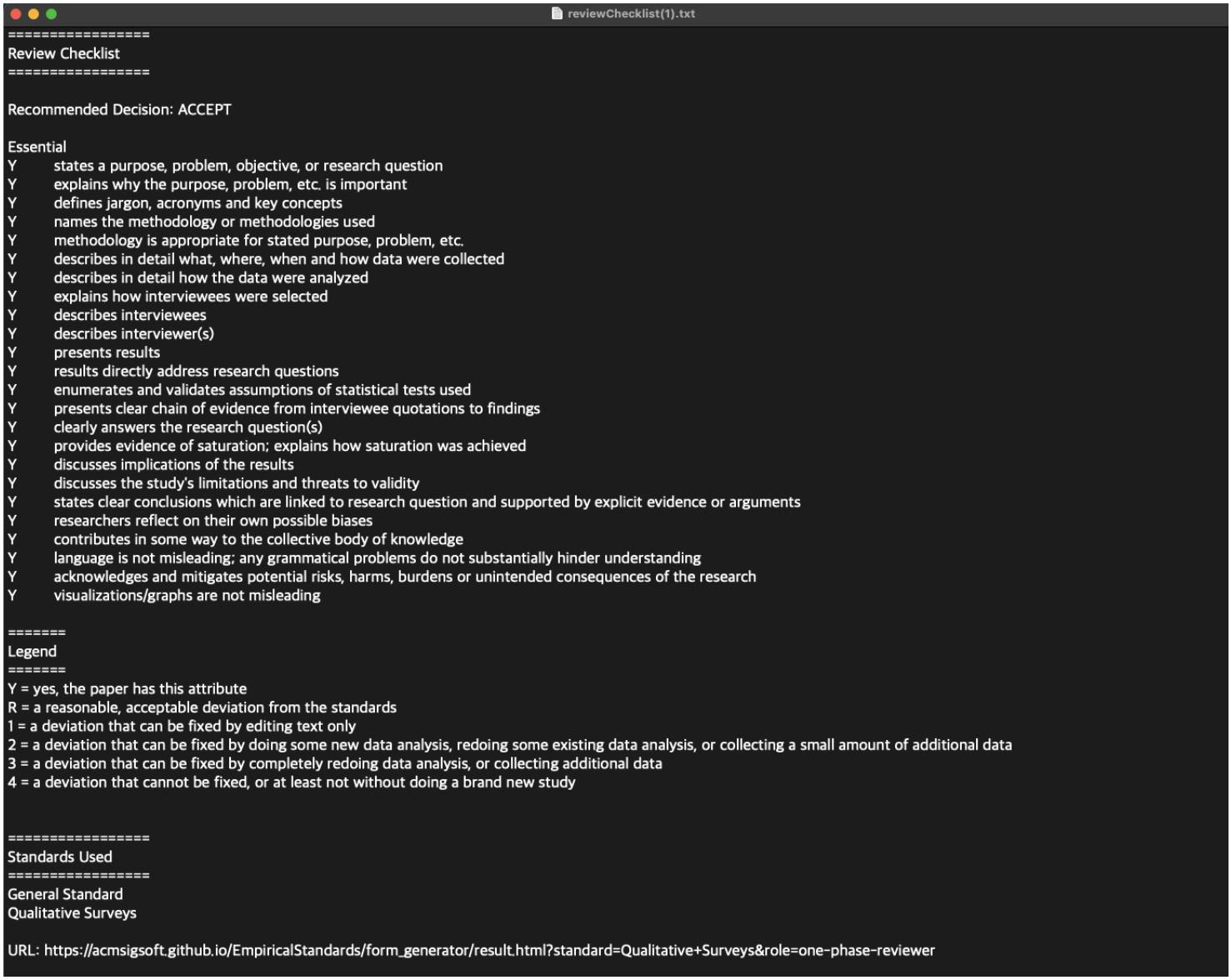


Figure 2.6: ESCG - Example of reviewChecklist.txt file for Qualitative Survey Checklist

Chapter 3

Methodology

My exploratory study aims to understand if the ESCs generated from the ESCG can be helpful to Ph.D. and Master's students when they review SE empirical research papers. This study involved conducting a Pilot and two Sessions of my study (Session 1 and Session 2). The timeline of my study is shown in Figure 3.1. The pilot and each session of my study were composed of a survey and a group discussion to obtain the participants' perceptions of the ESCs' comprehensibility, ease-of-use, and usefulness.

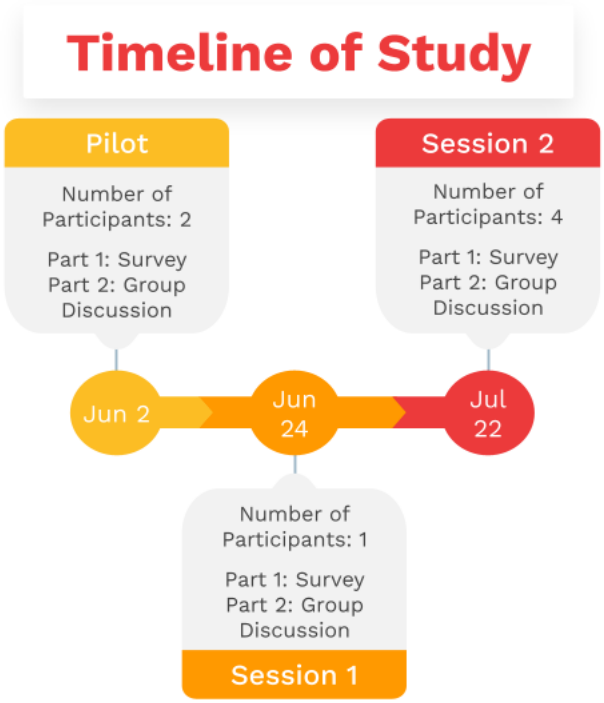


Figure 3.1: Timeline of Study: Pilot, Session 1 of Study, Session 2 of Study

3.1 Study Design (for Pilot, Session 1, and Session 2)

3.1.1 Survey Construction

Constructing the Comprehensibility section of the survey to answer RQ1

Molléri et al.'s paper, "Teaching students critical appraisal of scientific literature using checklists", was the foundation for the methodology of the pilot and sessions of my study. Molléri et al.'s methodology involved 76 students, where the students used two checklists to evaluate two research papers [13]. One paper had a case study methodology, and another had an experimental methodology. The authors then collected the students' opinions on using the checklists [13]. Based on Molléri et al.'s quantitative results, the students had positive opinions on the checklists' ease-of-use, and understandability [13]. In addition, some students stated they struggled more with reviewing the paper due to the presentation of the paper and their lack of experience in reviewing papers. However, several students highlighted room for improvement with the checklist by highlighting issues with the checklist items' structure, clarity, and terminology.

In the Appendix of Molléri et al.'s paper, there was a 5-point Likert scale survey question that asked their participants the following question"Please respond to what extent do you agree or disagree with the following statement: The questions were easy to understand. (strongly agree | ... | strongly disagree)"[13] Molléri et al.'s survey question was adapted for my survey question to say"Please say how much you agree or disagree with the following statements. The statements from the Empirical Standards Checklist for the first paper were overall easy to understand. (strongly agree | ... | strongly disagree)" to measure the participants' overall perceptions of the ESCs' comprehensibility.

The Appendix of Molléri et al.'s paper also contained the following survey questions "Please write the question numbers (from Table 1), if any, that were difficult to answer." and "What made it difficult to answer these questions?" [13]. These two survey questions were adapted for my survey questions to say"Do you understand all the statements from the 'Essentials' (ES) section of the first paper's Empirical Standards Checklist? (Yes — No)" and "If you answered 'No,' please list the ID with the corresponding reason why you do not understand the checklist statement. For example, 'ES#: I do not understand the statement because...'" to discover which checklist items from the ESCs were difficult for the participant to comprehend. A presentation of Molléri et al.'s survey questions and my survey questions regarding the participants' perceptions of comprehensibility can be shown in Figures 3.2 and 3.3.

C ADDITIONAL QUESTIONS REGARDING THE PERCEPTION OF USE OF THE CHECKLISTS

Please answer the following questions regarding the use of the given checklist:

A) Please respond to what extent do you agree or disagree with the following statement: Overall, the checklist was easy to use. (strongly agree | agree | neutral | disagree | strongly disagree)
B) Please respond to what extent do you agree or disagree with the following statement: The questions were easy to understand. (strongly agree | agree | neutral | disagree | strongly disagree)
C) Please write the question numbers (from Table 1), if any, that were difficult to understand.
D) Please respond to what extent do you agree or disagree with the following statement: The questions were easy to answer. (strongly agree | agree | neutral | disagree | strongly disagree)
E) Please write the question numbers (from Table 1), if any, that were difficult to answer.
F) What made it difficult to answer these questions?
G) Does the checklist cover all the important aspects (as mentioned in the guidelines for conducting case study research) for high quality case study research?

Figure 3.2: Molléri et al. - Perceptions of Comprehensibility Survey Questions

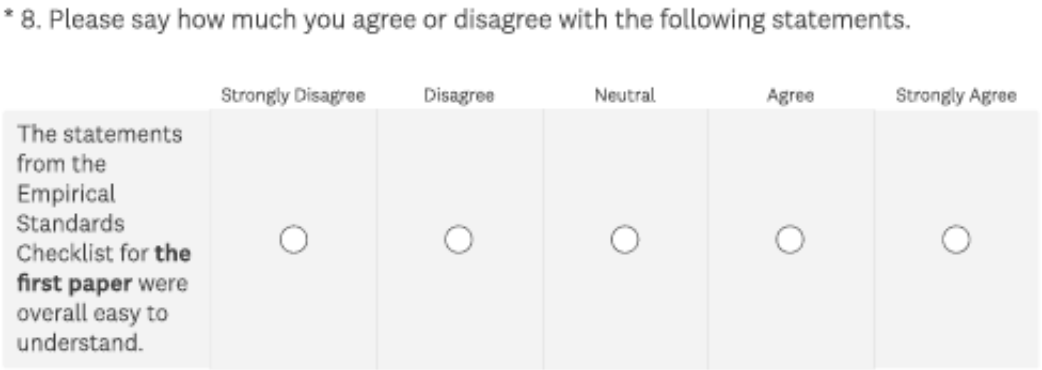


Figure 3.3: Section of Survey - Comprehensibility of the Empirical Standards Checklist for Reviewing Qualitative Survey Paper

Constructing the Ease-of-Use section of the survey to answer RQ2

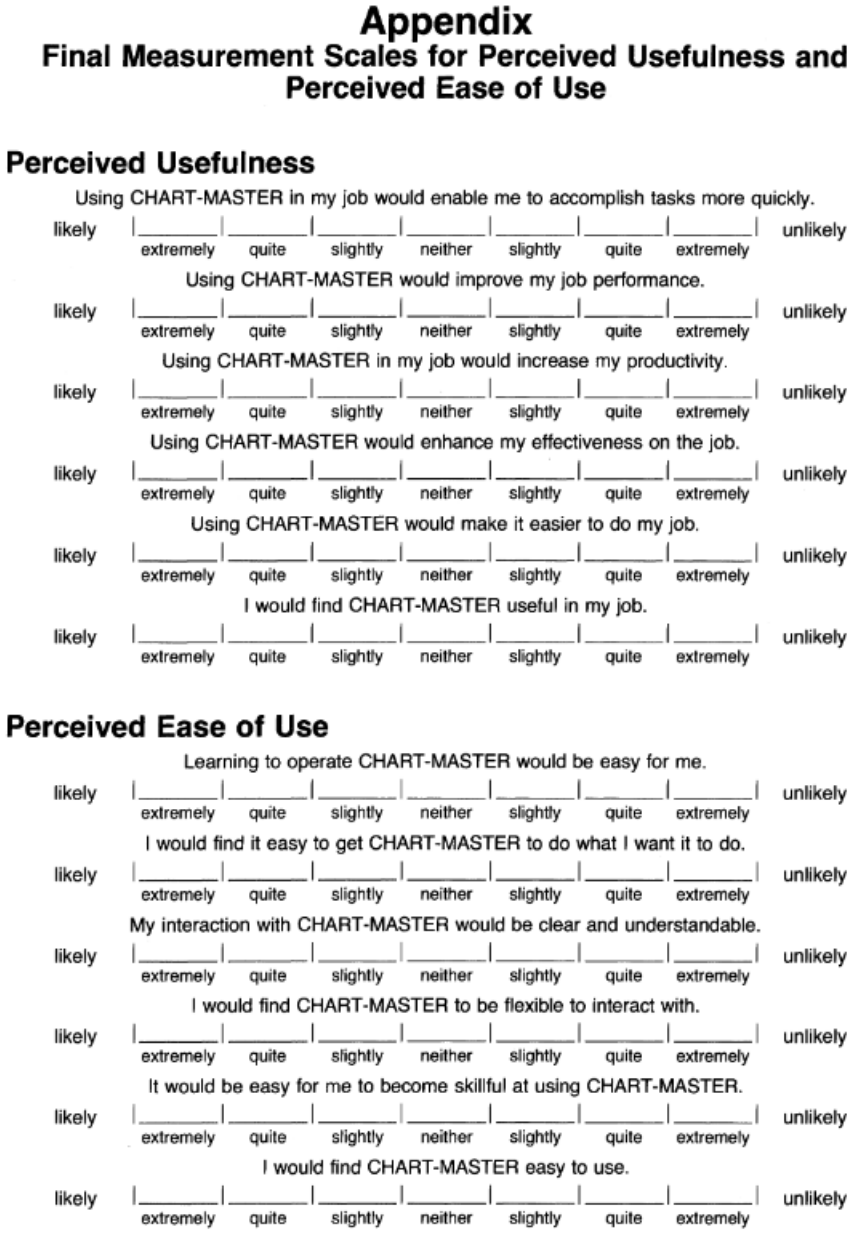
In addition, in the Appendix of Molléri et al.'s paper, there was a 5-point Likert scale survey question that asked their participants the following question "Please respond to what extent do you agree or disagree with the following statement: Overall, the checklist was easy to use. (strongly agree | ... | strongly disagree)" [13] Molléri et al.'s survey question was updated for my survey question to say "Please say how much you agree or

disagree with the following statements. Overall, I find the Empirical Standards Checklists easy to use. (strongly agree | ... | strongly disagree)" to measure the participants' overall perceptions of the ESCs' ease-of-use. A presentation of Molléri et al.'s survey questions and my survey questions regarding the participants' perceptions of ease-of-use can be shown in Figures 3.2 and 3.4.

Figure 3.4: Section of Survey – Ease-of-Use of Empirical Standards Checklist for Qualitative Survey Paper

Constructing the Usefulness section of the survey to answer RQ3

In "Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology," Davis provided an essential foundation for my survey questions to obtain the participants' perceptions of the ESCs' usefulness [1]. Davis aimed to create a scale to measure a user's perception of usefulness and ease-of-use, specifically for Information Technology (IT) systems. Davis built and refined this scale by performing two studies. The first study consisted of giving a questionnaire to 120 users (managers, administrative staff, programmers, etc.) in an IBM Toronto Development Laboratory to rate the usefulness and ease-of-use of two IT systems: "PROFS electronic mail" and "XEDIT file editor". After the first study, Davis refined the scales from ten items to six items for measuring the users' perceptions of ease-of-use and usefulness. Davis' then validated the scales once again in a second study. Davis recruited 40 MBA students from Boston University to evaluate two IBM PC graphics systems: "Chart-Master" and "Pendraw". Davis' final validated scale for measuring a user's perceptions of ease-of-use and usefulness for IT systems is illustrated in Figure 3.5. In addition, Davis' scale questions asking a user about their perceptions of an IT system's usefulness were adapted to my survey questions to ask participants their perceptions of the ESC's usefulness.



Appendix
Final Measurement Scales for Perceived Usefulness and Perceived Ease of Use

Perceived Usefulness

Using CHART-MASTER in my job would enable me to accomplish tasks more quickly.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

Using CHART-MASTER would improve my job performance.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

Using CHART-MASTER in my job would increase my productivity.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

Using CHART-MASTER would enhance my effectiveness on the job.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

Using CHART-MASTER would make it easier to do my job.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

I would find CHART-MASTER useful in my job.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

Perceived Ease of Use

Learning to operate CHART-MASTER would be easy for me.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

I would find it easy to get CHART-MASTER to do what I want it to do.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

My interaction with CHART-MASTER would be clear and understandable.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

I would find CHART-MASTER to be flexible to interact with.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

It would be easy for me to become skillful at using CHART-MASTER.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

I would find CHART-MASTER easy to use.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

Figure 3.5: Davis - Final Scale to Measure User's Perceptions of Usefulness and Ease-of-Use for Chart Master

Davis' 7-point Likert survey questions measuring a user's perception of usefulness were the following, "(1) I would find the system useful in my job (unlikely| ... | likely), (2) Using the system in my job would enable me to accomplish tasks more quickly (unlikely| ... | likely), (3) Using the system would make it easier to do my job (unlikely to likely), (4)

Using the system in my job would increase my productivity (unlikely| ... | likely), and (5) Using the system would make it easier to do my job (unlikely| ... | likely) [1]." I adapted Davis' survey questions to the following 5-point Likert survey questions "(1) Overall, the Empirical Standards Checklists would be useful to review research papers (strongly agree | ... | strongly disagree), (2) Using the Empirical Standards Checklists would enable me to review research papers more quickly (strongly agree | ... | strongly disagree), (3) Using the Empirical Standards Checklists would enable me to review research papers more easily (strongly agree | ... | strongly disagree), (4) Using the Empirical Standards Checklists would enable me to increase my understanding of research papers (strongly agree | ... | strongly disagree), and (5) Using the Empirical Standards Checklists would enable me to improve how I review research papers (strongly agree | ... | strongly disagree)." The reason that Davis' 7-point Likert scale survey questions are updated to 5-point Likert scale survey questions is to keep the scales for the survey questions consistent for the entirety of my pilot and sessions of my study. A presentation of Davis' survey questions and my survey questions regarding the participants' perceptions of usefulness can be shown in Figures 3.6 and 3.7.

PERCEIVED USEFULNESS		1	2	3	4	5	6	7		NA
1. Using the system in my job would enable me to accomplish tasks more quickly	unlikely	○	○	○	○	○	○	○	likely	○
2. Using the system would improve my job performance	unlikely	○	○	○	○	○	○	○	likely	○
3. Using the system in my job would increase my productity	unlikely	○	○	○	○	○	○	○	likely	○
4. Using the system would enhance my effectiveness on the job	unlikely	○	○	○	○	○	○	○	likely	○
5. Using the system would make it easier to do my job	unlikely	○	○	○	○	○	○	○	likely	○
6. I would find the system useful in my job	unlikely	○	○	○	○	○	○	○	likely	○

Figure 3.6: Davis - Perceptions of Usefulness Survey Questions

* 13. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Using the Empirical Standards Checklists would enable me to review research papers more quickly.	○	○	○	○	○
Using the Empirical Standards Checklists would enable me to review research papers more easily.	○	○	○	○	○
Using the Empirical Standards Checklists would enable me to increase my understanding of research papers.	○	○	○	○	○
Using the Empirical Standards Checklists would enable me to improve how I review research papers.	○	○	○	○	○
Overall, the Empirical Standards Checklists would be useful to review research papers.	○	○	○	○	○

Figure 3.7: Section of Survey - Usefulness of Empirical Standards Checklist for Reviewing Qualitative Survey Paper

3.1.2 Paper Selection

A Professor at the University of Victoria (UVic) provided a draft of the repository mining (RM) conference paper presented in Appendix A.6. A colleague in the Computer Human Interaction and Software Engineering Lab (CHISEL) at UVic provided a draft of the qualitative survey (QS) conference paper presented in Appendix A.5. This provided RM paper, and QS paper would be read and reviewed by the participants during the pilot and sessions of my study. I selected drafts of conference papers instead of finished conference papers to record more varied results from the participants as the papers were under development.

In addition, I selected a repository mining paper and a qualitative survey due to the research done by Storey et al. in "The who, what, how of software engineering research: a socio-technical framework [18]". In this paper, Storey et al. used their "Who-What-How framework" to analyze 151 papers from two different venues: the International Conference on Software Engineering (ICSE) and the Empirical Software Engineering Journal by Springer (EMSE). After analyzing the papers, Margaret-Anne Storey et al. found that the data strategy papers (59 EMSE papers and 58 ICSE papers) and sample survey strategy papers (16 EMSE papers and 5 ICSE papers) were the most prevalent papers out of the 151 papers. Due to these findings, I selected a repository mining paper and a qualitative survey paper for the participants to review during the pilot and sessions of my study.

3.1.3 Participant Selection

The Ph.D. and Master's students were chosen as my participants for the pilot and two sessions of my study because there is a higher chance that Ph.D. and Master's students have had some experience conducting research compared to undergraduate students' experience conducting research [5].

The participants were selected based on the following criteria:

1. Familiarity with the two empirical methodologies: repository mining and qualitative survey. These are the two methodologies within the research papers provided to the participants to review during the pilot and sessions of my study.

2. Had published one or more research papers.

3. Were either a Ph.D. or Master's student. Ph.D. students were preferred since they were expected to have more experience with research.

4. Located in North America. The Amazon gift cards provided to the participants were in Canadian or U.S. dollars, so the participants needed to be in North America.

3.1.4 Ethics

My ethics application was approved by the Human Research Ethics Board (HREB) for Session 1 of my study on March 16, 2022. My ethics application was approved by HREB for Session 2 of my study on July 8, 2022. The following documents for my ethics application are included in the Appendix.

1. Email Invitation to Participants in Session 1 of My Study (Appendix A.7)
2. Consent Form for Participants in Session 1 of My Study (Appendix A.8)
3. Email Invitation and Implied Consent Form for Participants in Session 2 of My Study (Appendix A.9)
4. HREB Application Approval for Session 1 of My Study (Appendix A.10)
5. HREB Application Approval for Session 2 of My Study (Appendix A.11)
6. TCPS2: Core - Certificate of Completion (Appendix A.12)

3.1.5 Recruitment for Pilot and Feedback from Pilot

Two participants from the Computer Human Interaction and Software Engineering Lab (CHISEL) at the University of Victoria were recruited for the pilot. During the pilot, some minor miscommunications arose, which were fixed for the subsequent sessions of my study. For example, participants did not know if they had to complete the entire survey during a specific time frame without interruption or if they had to wait for my direction to complete the survey. Another example is that the participants said they would be more comfortable knowing how much time was left to complete the survey. Thus, for the subsequent sessions of my study, they were asked to complete the survey without direction from me. In addition, a timer was shown to the participants to inform them how much time they had left to complete the survey. However, to be clear, the critical content for the survey and the group discussion did not change between the pilot and the sessions of my study. Only minor cosmetic changes were added to aid the participants. Due to an issue that arose during one of the sessions of my study, the data from the pilot was included in the overall data.

3.1.6 Recruitment for Sessions of My Study

The participants were Computer Science Master's and Ph.D. students recruited using two different strategies. Two recruitment strategies were needed due to the lack of interest from students to participate in the first session of my study (Session 1). Thus, a second recruitment strategy and a second session (Session 2) were required to obtain enough data for analysis.

Recruitment Strategy for Session 1 of My Study

For the first recruitment strategy, I used an email, which included an image advertising Session 1 of my study depicted in Figure 3.8, that was sent to various UVic mailing lists, the leaders of UVic Computer Science Research Groups, the leaders of UVic Student Clubs, and specific official UVic social media channels (the official UVic Reddit and the official UVic Engineering & Computer Science Discord). The monetary incentive for participation in Session 1 was a $20 Canadian Amazon Gift Card. However, using the first recruitment strategy, only six participants showed interest in Session 1 by filling out a Microsoft Form similar to the Microsoft Form included in Appendix A.4. An email was sent to the six participants with an attached consent form. If the participants signed and returned the consent form, the participants then agreed to participate in Session 1 and to be audio and video recorded for the entirety of Session 1. However, only one participant returned a signed consent form and attended Session 1 of my study on June 24, 2022.

Recruitment Strategy for Session 2 of My Study

Due to the lack of students' interest in participating in Session 1, I used a different recruitment strategy to recruit more participants. For this recruitment strategy, I increased the incentive to a $60 Canadian Amazon Gift Card or an equivalent valued Amazon Gift Card in $USD. Next, a tweet from the Twitter account called "@Cassandra_Cupryk" stated "Are you a Computer Science Master's or Ph.D. student who is interested in obtaining a $60 Amazon Gift Card by participating in a focus group? This focus group is a part of my Computer Science Master's Project. Look below for more info!" and presented an image advertising Session 2 of my study shown in Figure 3.9. The image included a link and a QR code to a Microsoft Form presented in Appendix A.4.

Any students interested in participating in Session 2 of my study filled out the Microsoft form with their contact information (first name, last name, and email) and answered the form's demographic questions. An email and a calendar invitation for Session 2's Zoom

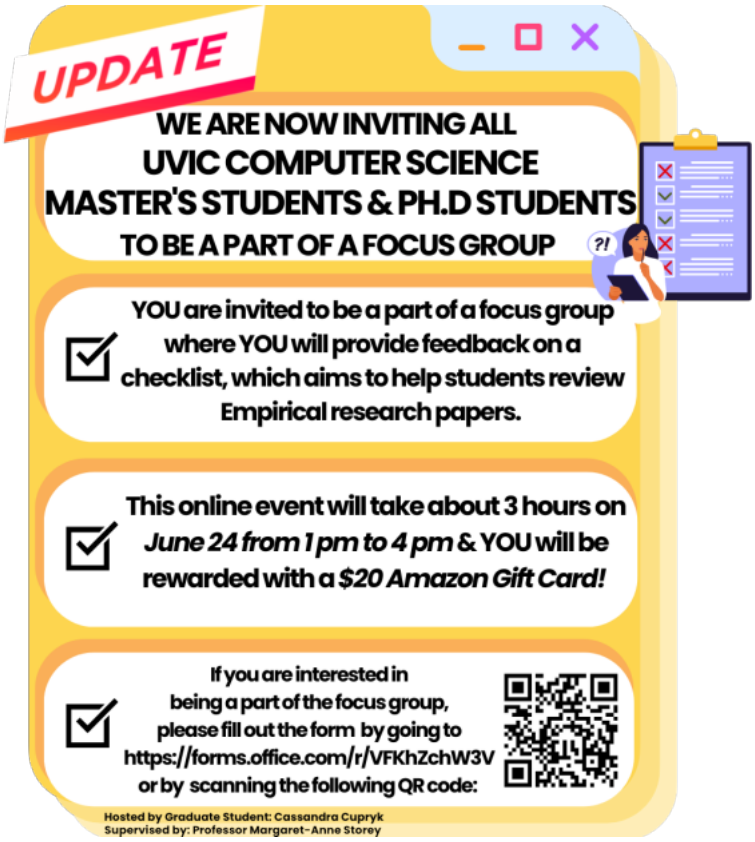


Figure 3.8: Recruitment Post for Session 1 of My Study

meeting were then sent to the students who fit the recruitment criteria for the session. In the email, the participants were informed that by attending Session 2 of my study on the outlined date and time, they implied their consent to participate in the session and to be audio and video recorded during the session. Using the second recruitment strategy, 64 students filled out the Microsoft Form to express interest in participating in Session 2. The email and Zoom meeting were then sent out to the five students who best fit the criteria chosen for recruiting participants outlined in section 3.1.3. In the end, four students attended Session 2 of my study on July 22, 2022.

3.2 Process for Pilot and Each Study Session

The pilot and each session of my study were split into two main parts. The first part involved the participants completing a survey hosted on "SurveyMonkey". The second part involved the participants having a group discussion.

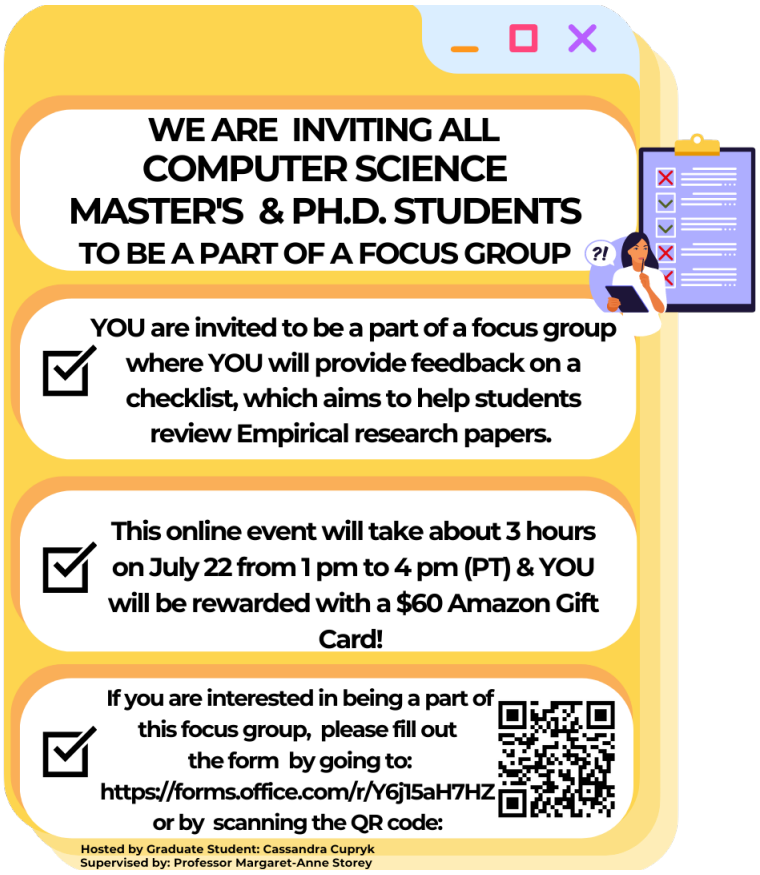


Figure 3.9: Recruitment Post for Session 2 of My Study

3.2.1 Part 1 of Pilot and Each Study Session: Survey

The survey was split into the following seven sections:

Section 1 of Survey: Read and Review Papers

For Section 1 of the survey, the participants were asked to read the two research papers (a QS paper and an RM paper) and then review these two research papers by using and completing the two corresponding ESCs. These two research papers were drafts of conference papers. The titles of both research papers were anonymized and did not include any identifying information about the authors.

Once the participant completed each ESC, the ESC would be available to download as a text file (.txt) named "reviewChecklist.txt" to their device. An example of what "reviewChecklist.txt" looked like is presented in Figure 3.10. The participant then copied and pasted the text of the "reviewChecklist.txt" file into the designated area of Section

1 of the survey. Participants were also asked to keep the windows of the two completed ESCs open to answer the questions in Sections 2 and 3 of the survey.

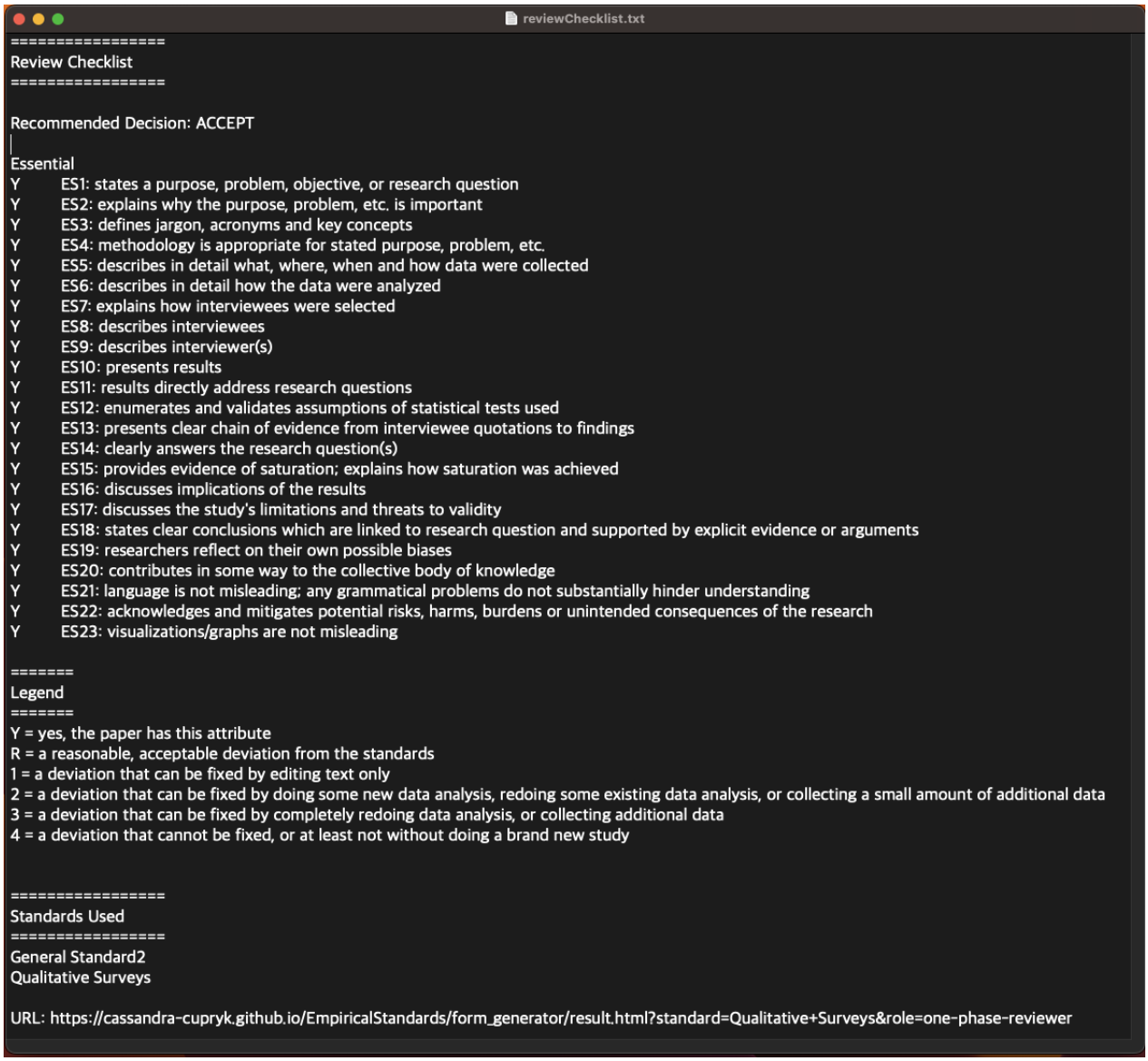


Figure 3.10: Example of reviewChecklist.txt file

Sections 2 and 3 of Survey: Understanding the Statements from the Empirical Standards Checklist for the Qualitative Survey Paper and the Repository Mining Paper

Sections 2 and 3 are intended to answer research question 1 (RQ1). Moreover, the survey questions for sections 2 and 3 involved asking questions about the participants'

opinions on the comprehensibility of the ESCs. These sections also asked the participants to highlight which checklist items were difficult for them to understand from the QS and the RM checklist.

Sections 4 and 5 of Survey: Opinions of Usefulness and Ease-of-Use

Sections 4 and 5 of the survey were used to answer research questions 2 (RQ2) and 3 (RQ3), respectively. The survey questions for sections 4 and 5 involved asking questions about the participants' opinions on the usefulness of the ESCs, as well as the ease-of-use of the ESCs.

Section 6 of Survey: Demographic Data Section

In the survey, the participant would select the empirical methodologies familiar to them, the number of papers they've published, whether they have taken a course on how to perform research, and whether they have used a tool or piece of software to review research papers.

Section 7 of Survey: Final Thoughts

In this section, the participant would provide their final thoughts about the ESCs that they could not share in the previous sections of the survey.

3.2.2 Part 2 of Pilot and Each Study Session: Group Discussion

For the second part of the pilot or each session of my study, there was a group discussion to collect feedback on their opinions on the pilot or session of the study. Moreover, the participants were asked to type an answer on the Zoom whiteboard to answer the question "What is your overall opinion of today's focus group?". However, some participants felt more comfortable answering the question via Zoom chat or audibly. Thus, all the participants' feedback was collected using the Zoom whiteboard, the Zoom chat, and the participants' audio.

3.3 After the Pilot and Sessions of My Study

After the two sessions of the study were over, the participants who attended the first and the second sessions of my study (not the pilot) were sent a $60 Canadian Amazon

gift card or an equally valued United States Amazon gift card.

3.4 Theme Code Analysis

The coding handbook was built iteratively while analyzing the qualitative data from the Pilot, Session 1, and Session 2. After the pilot or the session occurred, the audio recordings were transcribed with the aid of the dictation software provided by Microsoft Word. In addition, other qualitative data was obtained from the survey, the Zoom chat, and the Zoom whiteboard. All the qualitative data was then copied into Microsoft Excel. Then, the qualitative data was open-coded to answer RQ1, RQ2, and RQ3. Moreover, the open codes were primarily chosen to present the participants' perceptions of the ESCs' comprehensibility, ease-of-use, and usefulness. As the coding handbook was updated after each session, the open codes of the previous session's qualitative data were also updated. For example, new codes were obtained after analyzing the qualitative data of Session 1. Thus, the coding handbook and the codes for the Pilot's qualitative data were updated. Appendix A.1 provides the final table of theme codes derived from all the qualitative data.

The steps of my methodology have been explained above to show how the results will be collected to answer my three research questions. The next chapter will reveal the obtained results after conducting my study.

Chapter 4

Results

This chapter presents the data collected during my study. It is divided into the following sections: the participants' demographic data, the participants' decisions on whether to accept or to reject the Qualitative Survey (QS) paper or the Repository Mining (RM) paper, the participants' perceptions on the comprehensibility of the Empirical Standards Checklist (ESC) (RQ1), the participants' perceptions on the ease-of-use of the ESC (RQ2), and the participants' perceptions on the usefulness of the ESC (RQ3). To answer these research questions, I will present my results based on the collected quantitative and qualitative data, respectively.

4.1 Participants' Demographic Data

Table 4.1 presents the following demographic data of the participants: the number of papers that the participant has published, whether the participant has taken a course on how to perform research, and whether the participant has used a tool or a piece of software to review research papers.

Participant ID	Number of Papers that the Participant has Published	Whether the Participant has taken a course on how to perform research	Whether the Participant has used a tool or piece of software to review research papers
1	4	Yes	Yes
2	2	Yes	No
3	1	Yes	Yes

4	5+	Yes	No
5	4	Yes	No
6	0	Yes	Yes
7	2	Yes	Yes

Table 4.1: Demographic Data - Participants' Information

Figure 4.1 demonstrates that the participants were the most familiar with the following empirical standard methodologies: "Case Study", "Data Science", and "Experiment (with human participants)". However, none of the participants were familiar with the empirical standard methodology: "Meta-science".

Figure 4.2 highlights that **all** the participants had taken a university course intended to teach students how to perform research. Additionally, four participants (P1, P3, P6, P7) stated that they had used a tool or piece of software (other than the ESC) to review research papers.

In the survey, I asked the participants to provide the names of software or tools for reviewing research papers. One participant (P5) highlighted that they had used *"Robot"* to review research papers. However, I could not find this software or tool named "Robot" intended for reviewing research papers. P7 did not provide any names for tools or software for reviewing papers. Instead, they offered terms of statistical methods, *"I've used analytical tools like chi-square and regressional technique."*. P3 stated they used annotation tools to write notes to review papers, *"Annotation tools to write notes, but I am not sure if that's what you mean by'Software'."* Moreover, none of the participants provided names for alternative software or tools to the ESC for reviewing research papers.

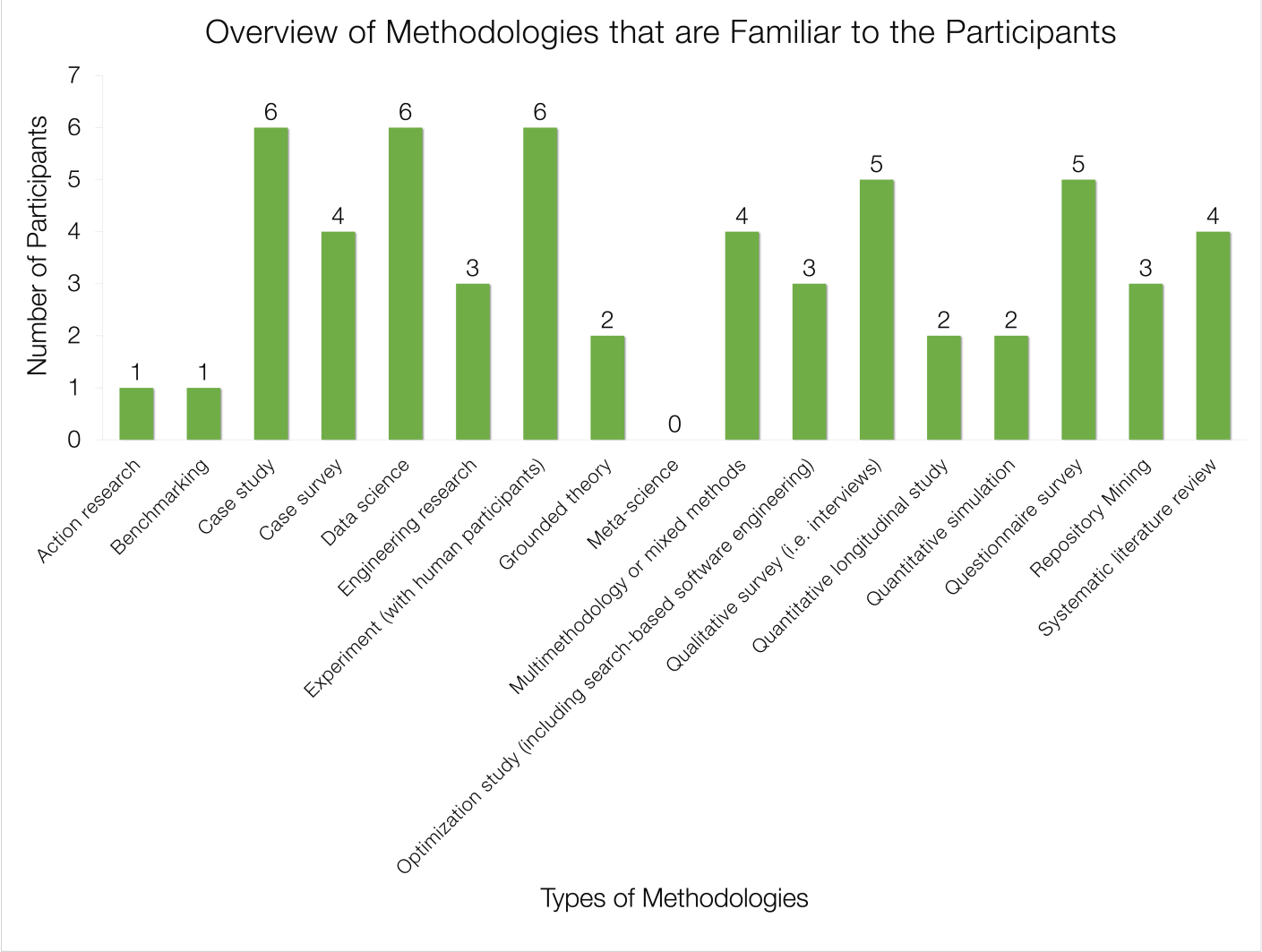


Figure 4.1: Demographic Data - Overview of the Empirical Standard Methodologies that are Familiar to the Participants

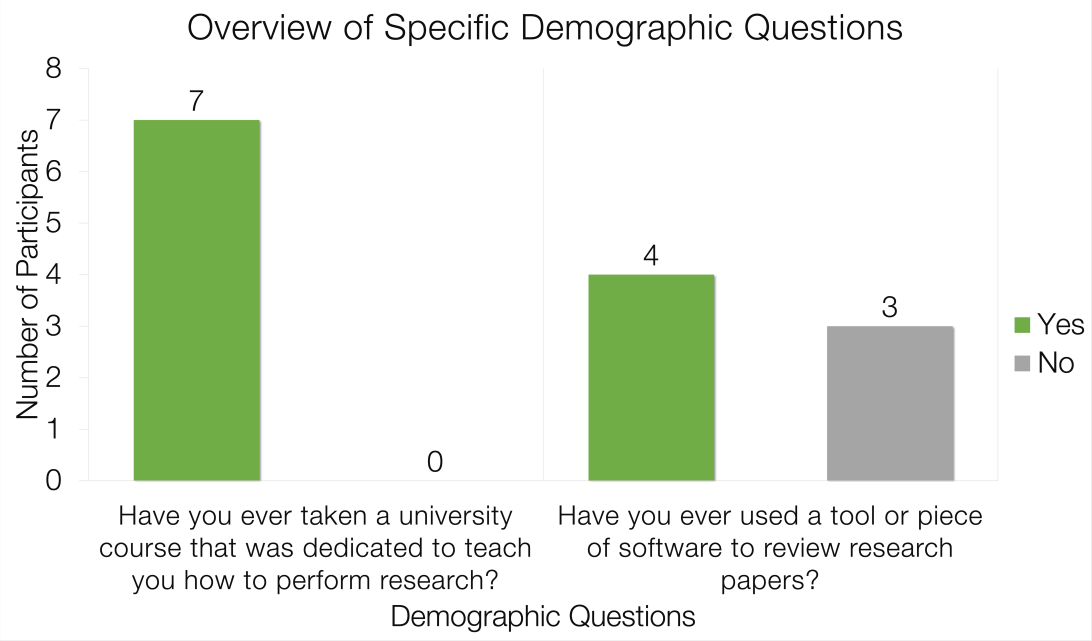


Figure 4.2: Demographic Data - Results of Demographic Survey Questions

4.2 Accepting or Rejecting Papers

In the first section of the survey, I asked the participants to use the ESC to help them decide if they should either reject or accept the QS paper and the RM paper. Figure 4.3 highlights that all five participants (P1, P2, P3, P4, P6) decided to accept the RM paper after reviewing the paper with the ESC and that two participants did not answer the survey question. In contrast, Figure 4.4 illustrates that four participants (P1, P3, P4, P7) rejected the QS paper after reviewing the paper with the ESC, and one participant did not answer the survey question. Three participants (P1, P3, P4) selected the same checklist item from the QS ESC as the reason for rejecting the QS paper: #15 in the "Essential" section of the QS ESC, **"ES15: provides evidence of saturation; explains how saturation was achieved".** Additionally, P7 selected all of the following checklist items from the essential section of the QS ESC as the reasons to reject the QS paper:

1. ES3: defines jargon, acronyms, and key concepts
2. ES4: explains why repository mining is appropriate for the proposed research problem
3. ES8: defines unit(s) of analysis or observation
4. ES12: describes the selected repositories
5. ES14: if the data obtained is too large to be processed in its entirety - explains why - explains the sampling strategy
6. ES16: describes data preprocessing steps
7. ES19: EITHER: uses previously validated measures, OR: assesses construct validity
8. ES23: enumerates and validates assumptions of statistical tests used
9. ES29: language is not misleading; any grammatical problems do not substantially hinder understanding

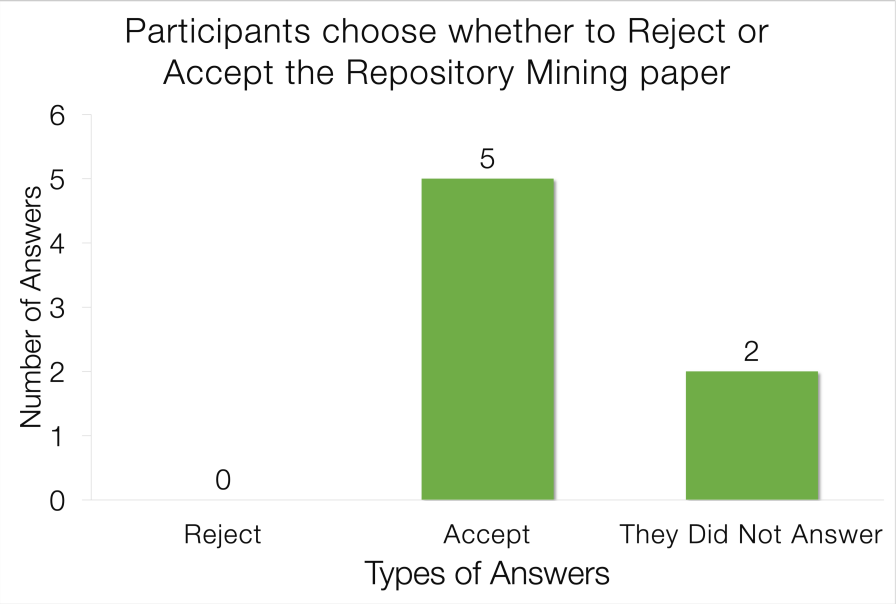


Figure 4.3: Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the Repository Mining Paper

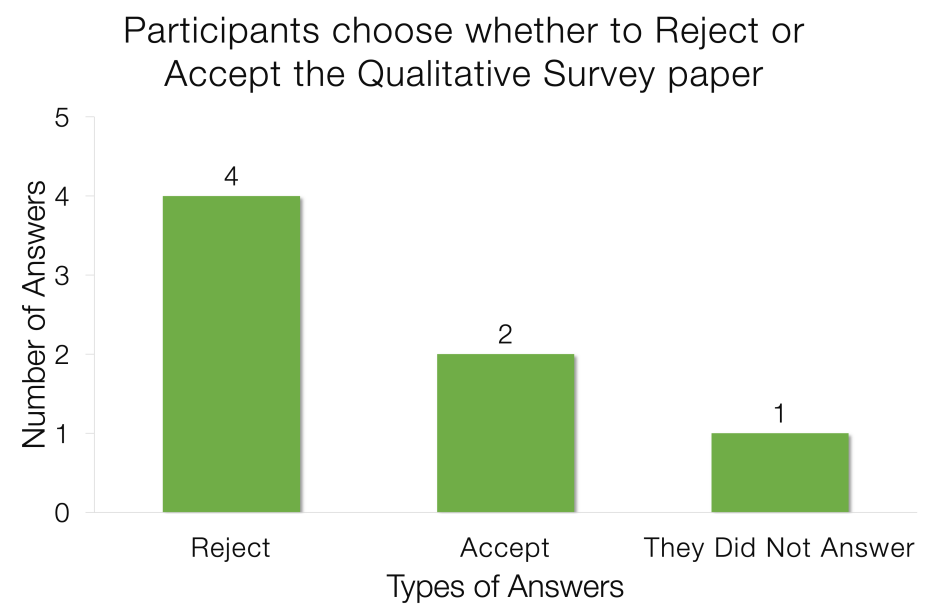


Figure 4.4: Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the Qualitative Survey Paper

4.3 RQ1: What are the Ph.D. and Master's students' perceptions concerning the comprehensibility of the Empirical Standards Checklist?

4.3.1 Participants' General Perceptions on the Comprehensibility of the Empirical Standards Checklists

Several participants voiced different perceptions of the comprehensibility of the ESCs. P3 had no difficulty with the comprehensibility of the ESCs saying *"I didn't have any questions on the checklists. They were pretty self-explanatory."* However, a couple of participants (P4, P5) commented that they needed help understanding some of the checklist items of the ESCs. P5 wrote that the ESCs forced them to do additional research to understand some of the checklist items, *"Made me look up many things."* P4 pointed out that inexperienced researchers may find the ESCs difficult to understand, *"Not for beginners in its current state."* Additionally, P4 also stated that they felt too inexperienced to understand the ESCs, *"I feel like I don't know enough to understand the checklists "* and that their inexperience is a barrier preventing them from using the ESCs, *"I would love to use it, but I need to learn **a lot** more about the things mentioned in the survey."*

4.3.2 Participants' Perceptions on the Comprehensibility of the Qualitative Survey Empirical Standards Checklist

In Figure 4.5, I present the seven participants' overall perceptions on whether the checklist items from the QS ESC were easy to understand. Our results indicate that **five out of the seven participants (P1, P2, P3, P5, P6) agreed** that the statements from the QS ESC were overall easy to understand.

Figure 4.6 presents a more detailed view of which sections (Essential (ES), Desirable (DE), Extraordinary (EX)) of the QS ESC were challenging to understand for the participants. As indicated in Figure 4.6, **five out of seven participants(P2, P3, P5, P6, P7)** were able to understand all the statements in the Essential section of the ESC, **six out of seven participants (P1, P2, P3, P5, P6, P7)** were able to understand all the statements in the Desirable section of the ESC, and **all seven participants** were able to understand all the statements in the Extraordinary section of the ESC.

During the survey, I asked the participants to identify the checklist items in each

section (Essential, Desirable, Extraordinary) that were difficult for them to understand. For the Essential section, P4 revealed that they did not have enough previous knowledge to understand the term "saturation" in checklist item #ES15 (#ES15: provides evidence of saturation; explains how saturation was achieved) stating *"ES15: "saturation" is a new concept for me; I had to look it up and have a shaky understanding of it."* In addition, P1 commented that there were some comprehensibility issues regarding the meaning and subjectivity of terms in the same checklist item #ES15, *"Yes #15. What do you mean by saturation? Or what do you mean by reasonable? The checklist has these vague and subjective words. "*

For the Desirable section, P4 had difficulty understanding checklist item #DE18 (#DE18: validates results using member checking, dialogical interviewing, feedback from non-participant practitioners or research audits of coding by advisors or other researchers) by stating *"I'm not familiar with these methods of result validation, and I'd have to read the paper to brush up on them."* P4 also pointed out another checklist item #DE1 (DE1: states epistemological stance) that they had trouble understanding and stated that they did not have enough expertise to identify epistemological stances in a paper, *"DE1: I'm not familiar with different epistemological stances to the point where I'm comfortable identifying them."*

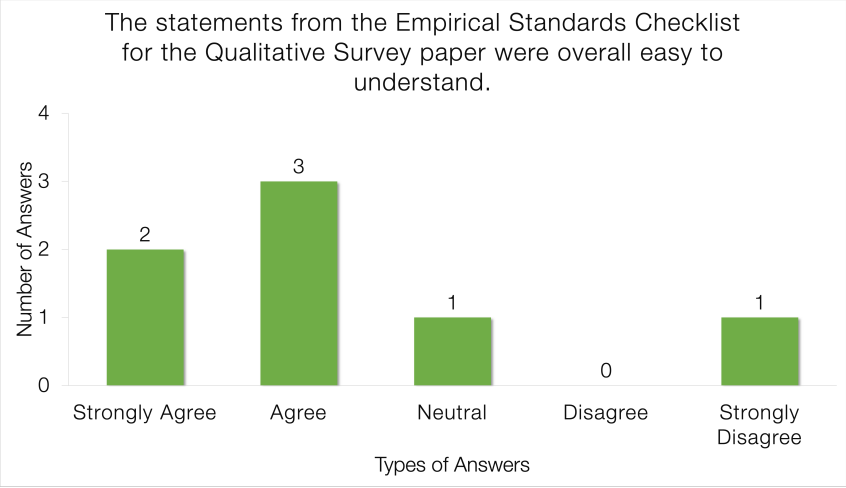
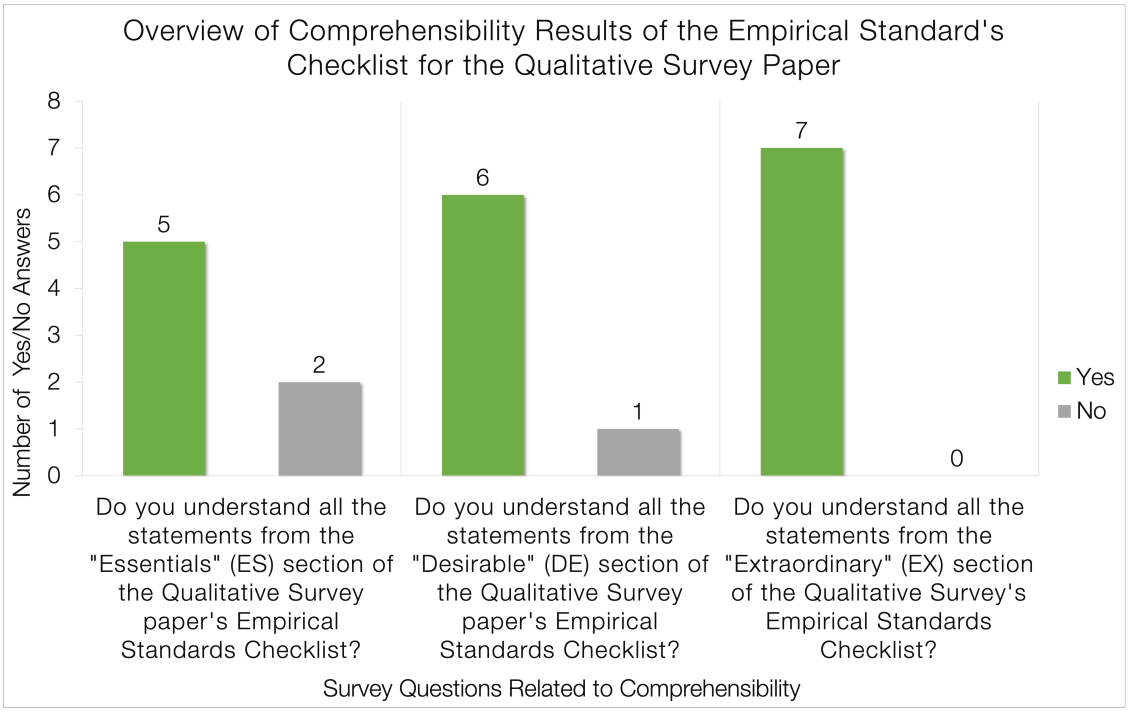


Figure 4.5: Comprehensibility - Results for the Survey Question: "The Statements from the Empirical Standards Checklist for the Qualitative Survey Paper were Overall Easy to Understand"



(a) Comprehensibility of the Essential section of the checklist for the Qualitative Survey Paper

(b) Comprehensibility of the Desirable section of the checklist for the Qualitative Survey Paper

(c) Comprehensibility of the Extraordinary section of the checklist for the Qualitative Survey Paper

Figure 4.6: Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the Qualitative Survey Paper

4.3.3 Participants' Perceptions on the Comprehensibility of the Repository Mining Empirical Standards Checklist

In Figure 4.7, I present the seven participants' overall perceptions on whether the checklist items from the RM ESC were easy to understand. Moreover, **five out of seven of the participants (P1, P2, P3, P5, P6) agreed** that the statements from the RM ESC were overall easy to understand.

Figure 4.8 outlines a more detailed presentation on which sections (Essential (ES), Desirable (DE), Extraordinary(EX)) of the RM ESC were difficult for the participants to understand. Five out of seven participants (P1, P2, P5, P6, P7) were able to understand all the statements in the "Essential" section of the ESC, six out of seven participants(P1, P2, P3, P5, P6, P7) were able to understand all the statements in the "Desirable" section of the ESC, and six out of the seven participants (P1, P2, P3, P4, P5, P6) were able to understand all the statements in the "Extraordinary" section of the ESC.

In the survey, I asked the participants to pinpoint the checklist items in each section (Essential, Desirable, Extraordinary) that were difficult for them to understand. For example, P4 felt overwhelmed by the amount of information in checklist item #ES20 (ES20: if predictive modeling is used, complies with the Data Science Standard) and stated *"ES20: The data science standard is A LOT to take in!"* Similar to the comments of P4 in Section 4.3.2, P4 did not understand the term "epistemological stance" in the RM ESC checklist item #DE1 (DE1: states epistemological stance)), stating *"DE1: Again, epistemological stance."*

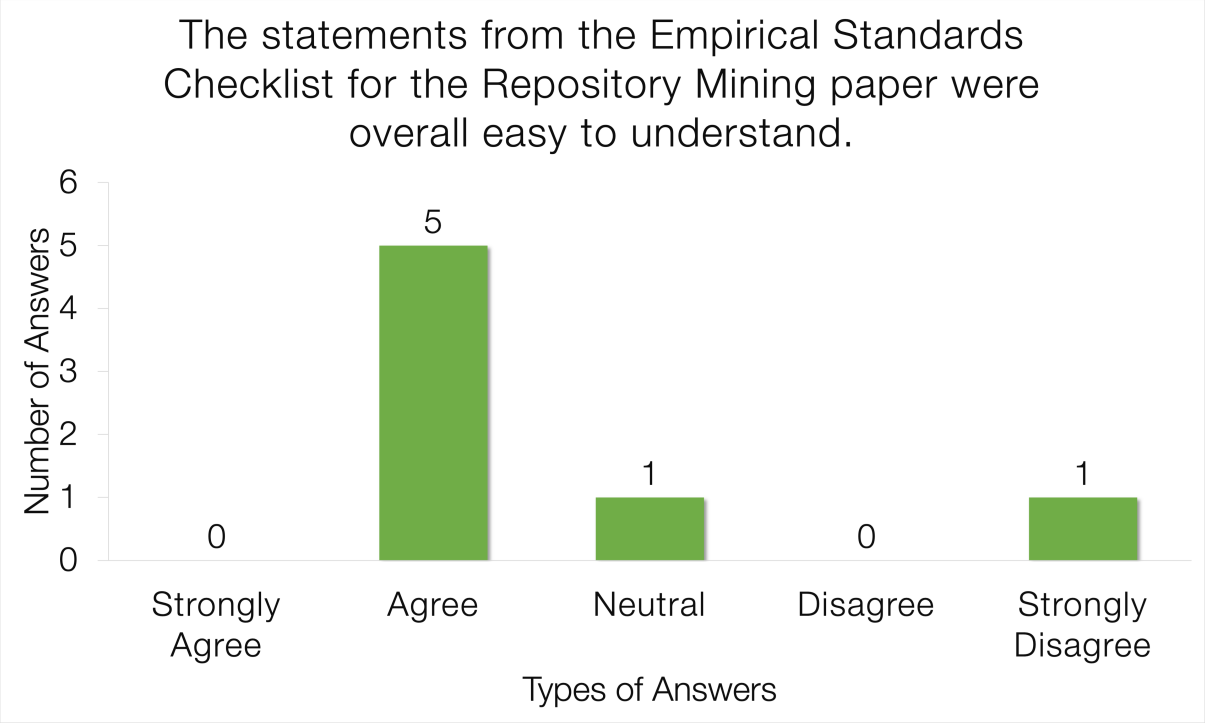
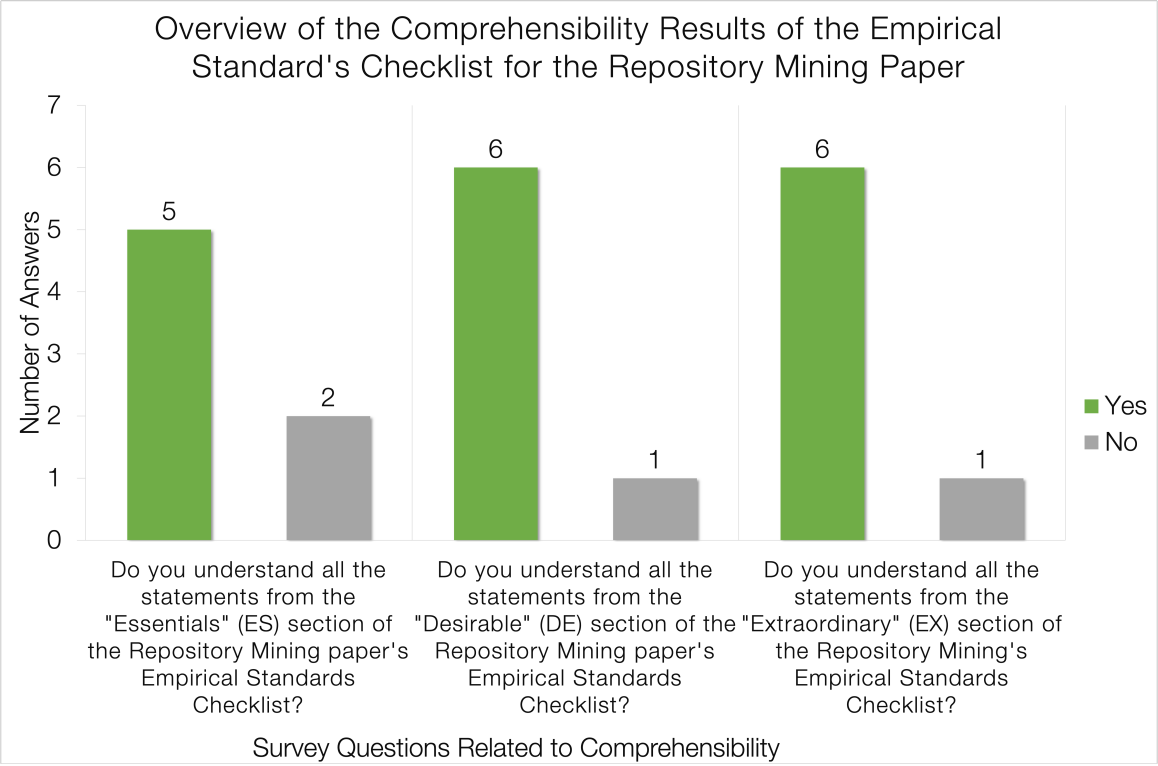


Figure 4.7: Comprehensibility - Results for the Survey Question: "The Statements from the Empirical Standards Checklist for the Repository Mining Paper were Overall Easy to Understand"



(a) Comprehensibility of Essential section of the checklist for the Repository Mining Paper

(b) Comprehensibility of Desirable section of the checklist for the Repository Mining Paper

(c) Comprehensibility of Extraordinary section of the checklist for the Repository Mining Paper

Figure 4.8: Comprehensibility - Overview of Quantitative Comprehensibility Results of the Empirical Standards Checklist for the RM Paper

4.4 RQ2: What are the Ph.D. and Master's students' perceptions concerning the ease-of-use of the Empirical Standards Checklist?

Figure 4.9 shows that **four out of the seven participants (P1, P2, P3, P6) agreed** that the ESC was easy to use, while two participants (P4, P7) strongly disagreed.

Even though the majority of participants agreed that the ESC was easy to use, several participants voiced their critiques on the ease-of-use of the ESC. P2 highlighted that if they accidentally closed the window hosting the ESC, they would then lose all their progress when completing the ESC, *"I read the paper and reviewed the paper, and then, unfortunately I closed my window."* P2 also voiced that they had difficulty reading the

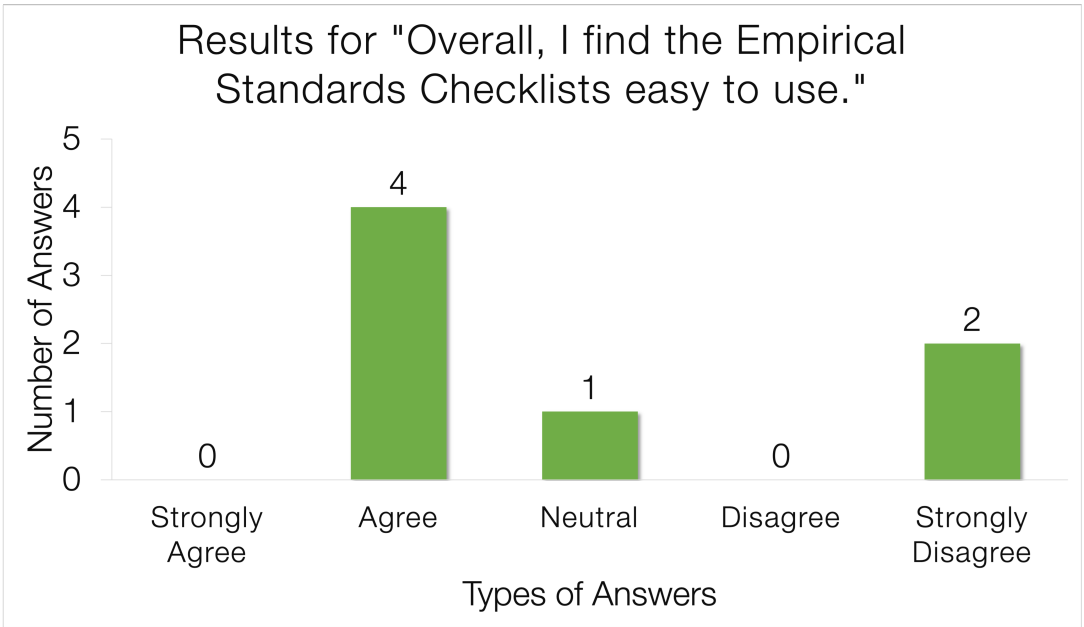


Figure 4.9: Ease-of-Use - Results for the Survey Question: "Overall, I find the Empirical Standards Checklists Easy to Use"

text of the ESC since the text was situated too close together, *"One thing that would be nice is a way to make it easier to read since sometimes the text is bunched up with the items above and below the one you are reading."*

P1 pointed out that the ESC's "Desired" and "Extraordinary" sections were hidden from them until the "Essential" section of the ESC is completed, *"Furthermore, why not allow the reviewer to consider the Desired and Extraordinary items even if the paper receives some "Not reasonable" mark during the Essential list?"*

P1 also stated that when they had hovered their mouse over underlined checklist items, the ESC would provide more information on those checklist items. However, P1 pointed out that using an underline is not a familiar way of providing additional information to a user. Instead, P1 recommended that the checklist items had a more information icon button similar to ℹ, which usually signifies that there is more information to a user, *"When you have something that you want to explain to us, it's underlined. It took me some time to understand that if I mouse over a checklist item, I would be seeing details of what that checklist item means. Usually, we have an icon sign like ℹ. I know that I will see additional information with this icon. Currently, it's not following the pattern that I would be expecting."*

4.5 RQ3: What are the Ph.D. and Master's students' perceptions concerning the usefulness of the Empirical Standards Checklist?

To obtain the overall sentiment of the participants on the usefulness of the ESCs, I asked them to express their perceptions on the statement, "Overall, the ESCs would be useful to review research papers." Figure 4.10 shows that**five out of seven of the participants (P1, P2, P3, P4, P6) agreed** with that statement. In addition, some participants (P2 and P4) stated that the ESCs were very useful for reviewing research papers. For example, in the words of P4, *"It's super useful."* Moreover, P2 emphasized, *"The checklists are a useful and systematic way to review papers."*

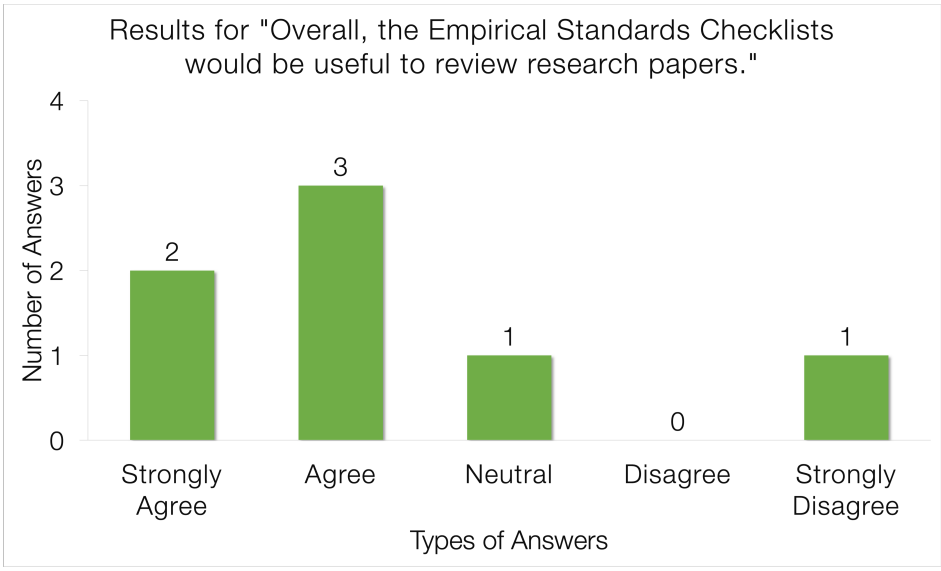


Figure 4.10: Usefulness - Overview of Quantitative Usefulness Results of the Empirical Standards Checklist for the Qualitative Survey Paper

I examined the participants' perceptions of the usefulness of the ESC in terms of how the tool enables them to: (1) review research papers more quickly (Section 4.5.1), (2) review research papers more easily (Section 4.5.2), (3) increase their understanding of research papers (Section 4.5.3), and (4) improve how they review research papers (Section 4.5.4).

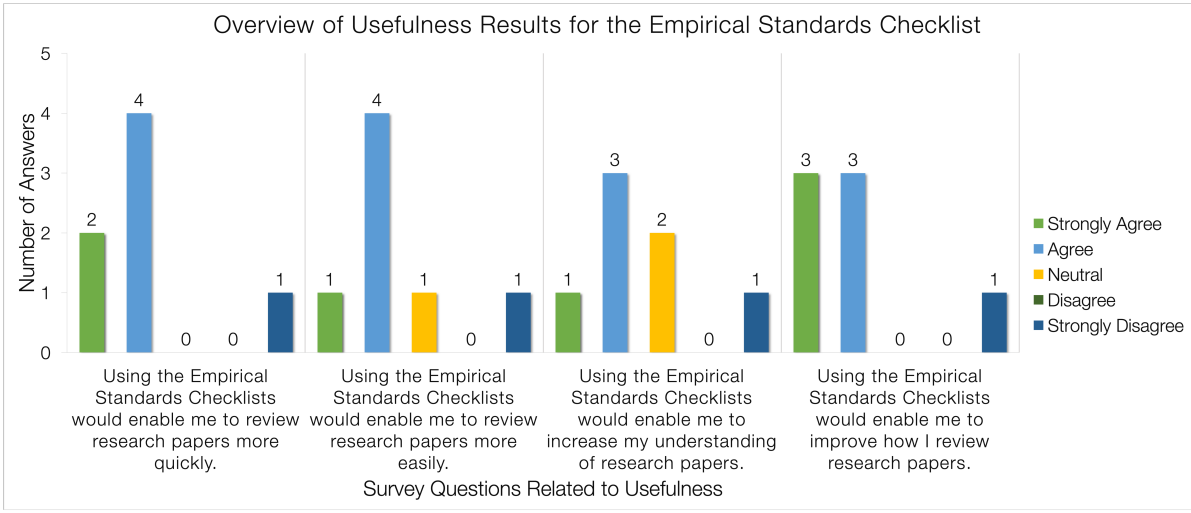


Figure 4.11: Usefulness - Overview of Quantitative Usefulness Results of the Empirical Standards Checklist for the Qualitative Survey Paper

4.5.1 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to review research papers more quickly?

Based on my survey, **six out of the seven participants agreed (P1, P2, P3, P4, P5, P6)** that using the ESC would enable them to review research papers more quickly. However, one participant (P7) strongly disagreed. The ESC helps Ph.D. and Master's students speed up the reviewing process by narrowing the criteria reviewers need to evaluate a research paper.

4.5.2 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to review research papers more easily?

Based on my survey, Figure 4.11 shows that **five out of the seven participants (P1, P2, P4, P5, P6) agreed** with the sentiment "Using the Empirical Standards Checklists would enable me to review research papers more easily." In contrast, one participant (P7) strongly disagreed. However, during the group discussion, P7 pointed out *"the ESC helped me learn an easier way of analyzing research papers."*

4.5.3 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to increase their understanding of research papers?

Based on my survey, **four out of seven participants (P1, P3, P4, P6) agreed** that using the ESCs would enable them to increase their understanding of research papers. Two participants (P2 and P5) were neutral, and one participant (P7) strongly disagreed. During the group discussion, some participants highlighted that the ESC would be useful for improving their research papers, for instance, in words of P3, *"Now I can focus on fixing and getting everything inside of the checklist,"* and *"Now I am aware of the considerations in the extraordinary category to improve my work."*

4.5.4 To what extent do Ph.D. and Master's students agree that the Empirical Standards Checklists would enable them to improve how they review research papers?

Based on my survey, **six out of the seven participants (P1, P2, P3, P4, P5, P6) agreed** that using the ESC would enable them to improve how they review research papers, and only one participant (P7) strongly disagreed. Additionally, from the group discussion, P2 stated that the use of the checklist would reduce the possibility of the reviewer introducing their bias when reviewing research papers or in the words of P2, *"I believe that a systematic way to review papers is good because it alleviates problems such as how the reviewer might be feeling that day and their own biases into how the research should have been conducted."*

Even though most participants agreed that the ESC would enable them to improve how they review research papers, a couple of participants voiced concerns about how the ESC could influence their decision to reject or accept a paper. For instance, P1 stated *"If I wasn't an experienced person, I might be influenced by this project, but I still see the value of what it's doing. It should help me evaluate all these items, but it should not say whether the paper should be accepted or rejected."* Furthermore, P1 proposed that the ESC should highlight the issues based on the ESC instead of telling the participant to accept or reject the paper to minimize bias. In this case, allow the reviewers to decide if the papers satisfy their quality criteria. For example, in the words of P1, *"There should be some red flag. Something that would indicate an improvement opportunity. That doesn't mean I would reject the paper if one of the items, even if they are considered essential,*

does not satisfy my perspective."

In addition, P2 commented that even though researchers have validated the ESC, there is still a concern that a reviewer may use the ESC as a tool for rejecting a paper. For example, P2 said *"I know they're validated, but it's still like if somebody were to find this website and use that as their sole reason for rejecting a paper."*

Another example is that P1 felt that the ESC was missing two checklist items and a third response for each checklist item. For instance, P1 stated that there was a missing checklist item that would evaluate the format of the paper, *"I missed something related to the paper's formatting or editorial stuff like missing references, problems with using italic, other quotations, and things that were not following the pattern that I would be expecting for, high-level standards of writing. If I were a reviewer, I would like to have this because it's a condition for the paper to be approved."* In addition, P1 said that they wanted a checklist item that would evaluate the paper's grammar, stating *"there wasn't one item related to grammar."* Finally, P1 suggested that the ESC should include a third option in the response considering "Not Applicable" and additional considerations for evaluating the format of the paper. For instance, P1 stated *"I would like to have the option to say not applicable."*

Chapter 5

Discussion

In this chapter, I summarize and also discuss an implication that can aid Ph.D. and Master's students in the future as well as highlight recommendations for the user interface (UI) designers and the research community working on the ESC. I present some new insights that emerged that were not directly related to the research questions I posed in my study. In addition, I present threats to the validity and limitations of my study.

5.1 An Implication for Ph.D. and Master's students

There is one implication for Ph.D. and Master's students regarding the future use of the ESC. The ESC does not only have to be used by Ph.D. and Master's students to review papers but can also be used by Ph.D. and Master's students to write their papers. In section, 4.5.3, a participant stated that they would want to use the ESC to improve how they wrote their papers. They even said they wanted to ensure their papers satisfied the checklist items from the "Extraordinary" section of the ESC.

5.2 Recommendations for the Empirical Standards Checklist

From my study, several recommendations emerged for the UI designers and the research community of the ESC. The recommendations for the UI designers are the following: improve the readability of the ESC, make the Desirable and Extraordinary checklist items visible at all times, abide by the user experience design (UX) standards, and allow the user to save their progress when completing the ESC. The recommendations for the research

community include: improving the comprehensibility of the ESC, building a glossary for confusing and complex terms, and training users on how to use the ESC. I discuss these insights in more detail below.

5.2.1 Recommendations for the UI Designers of the Empirical Standards Checklists

Improving the readability of the ESC. Participants recommended improving the readability of the ESC. A participant in Section 4.4 stated that the text of the checklist items should be more spaced out from the text above and below.

Enhancing the visibility of Desirable and Extraordinary items in the ESC. In section 4.4, a participant commented on the fact that certain sections ("Desirable" and "Extraordinary") of the ESC are not always visible to them until they've completed the checklist. Moreover, another recommendation would be to have all sections of the ESC visible at all times rather than have certain sections hidden.

Using general UX standards. Another recommendation for the designers of the ESCs would be to improve the website based on the general UX standards of websites. In section 4.4, a participant discussed how a checklist item should not have been underlined to signify that there is more information. Instead, an icon similar to this information icon: ℹ should be used to indicate that there is more information.

Adding a "Save progress" option. In section 4.4, a participant mentioned an issue they ran into while completing the ESCs. At one point during the study, the participant had accidentally navigated to another page while completing the ESC, and all the checklist items they had already completed disappeared. Moreover, this issue would be prevented from arising again if the data of the ESC could be saved.

5.2.2 Recommendations for the Research Community of Empirical Standards Checklist

Improving the comprehensibility of the ESC. The current version of the ESC may not be usable by all Ph.D. and Master's students since many participants pointed out comprehensibility barriers that prevented them from reviewing the papers. In section 4.3.1, one participant said that the ESC is not ready for beginner researchers to use in its current version. In section 4.3.2, several participants highlighted checklist items that prevented them from completing the review of the QS paper. In section 4.3.3, a participant

mentioned a checklist item containing an overwhelming amount of information.

Facilitating a glossary of terms to increase the comprehensibility of specific terms. In section 4.3.2, a few participants struggled with specific terms in the checklist. Moreover, one recommendation would be to provide additional information and clarity on these specific terms, so the comprehensibility of the ESC could be improved.

Training users on how to use the ESC. Another recommendation is to figure out a way to train new users on how to use the ESC. For example, there could be a quick tutorial to educate new users on how to use the ESC.

5.3 Validity threats and limitations

I highlighted the threats of validity following the criterion from Korstjens et al. Guba, which was divided into internal validity, construct validity, and reliability [4] [10].

5.3.1 Internal Validity

One potential threat was the time constraint for the reviewing paper tasks. Participants may have been influenced by the time constraint while reviewing the papers. However, I did run a pilot study to determine whether the time to read and review the papers was enough for the participants.

5.3.2 Construct validity

The design of the group discussion for the study may have influenced the participants' opinions. All participants were together when discussing their views on the ESC. However, I played the role of the moderator for the group discussion, ensuring all participants expressed their honest feedback and justified the rationales behind their responses.

5.3.3 Reliability

Another potential threat was that I was the only person who collected and analyzed the data. Therefore, I could have introduced bias since only one person ran and coded the data. However, I had several discussions with my supervisors to ensure the credibility of my study's results.

Chapter 6

Conclusion

I conducted an exploratory study with seven participants, consisting of Master's and Ph.D. students, to collect their perceptions on the Empirical Standards Checklists's comprehensibility, ease-of-use, and usefulness. During the exploratory study, I observed the students complete a survey using the Empirical Standards Checklist to read and review the Qualitative Survey and Repository Mining papers. In addition, I had a group discussion about collecting their opinions on how the study was conducted. Most participants agreed that the Empirical Standards Checklists were easy to understand, easy to use, and useful. However, the participants provided some feedback regarding the content of the Empirical Standards Checklists and the design of the Empirical Standards Checklists. For instance, there were some clarity issues regarding some terms in the Empirical Standards Checklists' checklist items. Another instance was a design issue that could cause the user to lose their progress when filling out the checklist if they navigated to another window. Future work could investigate whether using the guidelines would help researchers write better papers. Another idea is determining how to facilitate the adoption of the Empirical Standards Checklists into the Software Engineering research community by considering the researchers' tools and workflows. One last idea is to investigate the usability of the Empirical Standards Checklists. I hope that the insights from my project will help other researchers perform more studies evaluating the Empirical Standards Checklists to improve the Empirical Standards Checklists.

Appendix A

Additional Information

A.1 Final Theme Codes

Table A.1 presents the theme codes that are needed to evaluate the qualitative data from the survey, the group discussion's whiteboard and chat, and the audio transcript during the entire study.

Open Code	Definition
Checklist Alternatives	Participants provided software and tool alternatives that they've used to review research papers.
Checklist Comprehension	Participants expressing opinions on the comprehensibility of the checklist.
Checklist Design	Participants expressing opinions on the design of the checklist.
Checklist Ease-of-Use	Participants expressing opinions on the usability of the checklist.
Checklist Usefulness	Participants expressing opinions on the usefulness of the checklist.
Checklist Uses	Participants expressing opinions on the uses of the checklist.
Paper Choice	Participants expressing opinions on the research papers chosen for the study.
Paper Familiarity	Participants expressing opinions on how the research papers were familiar to them.
Study Feelings	Participants expressing opinions on how the study was performed.
Study Obstacles	Participants expressing opinions on any obstacles the participants faced during the study.
Study Tools	Participants expressing opinions on the tools used to facilitate the study.

Table A.1: Theme Codes for the Qualitative Data Analysis.

A.2 Survey for Pilot Study

Survey: Introducing the Empirical Standards Checklist

1. Section 1: Read and Review the 2 Research Papers

Please be aware that the time limit to complete Section 1 is 2 hours.

* 1. The time limit is **1 hour** to review **the first paper.**

- ☐ Download **the first paper** from Microsoft OneDrive by clicking Here.
- ☐ Read **the first paper.**
- ☐ Review **the first paper** by completing the checklist that is listed Here.
- ☐ To complete **section 3** of the survey, it is important to keep the window of the checklist of **the first paper** open.
- ☐ After completing the checklist for **the first paper**, do not forget to download the text file titled **"reviewChecklist.txt"** by clicking the Download button at the bottom of the window.

* 2. Please copy and paste all the text from your downloaded text file titled **"reviewChecklist.txt"** for **the first paper** into the text box below.

* 3. The time limit is **1 hour** to review **the second paper.**

- ☐ Download **the second paper** from Microsoft OneDrive by clicking Here.
- ☐ Read **the second paper.**
- ☐ Review **the second paper** by completing the checklist that is listed Here.
- ☐ To complete **section 4** of the survey, it is important to keep the window of the checklist of **the second paper** open.
- ☐ After completing the checklist for **the second paper**, do not forget to download the text file titled **"reviewChecklist.txt"** by clicking the Download button at the bottom of the window.

* 4. Please copy and paste all the text from your downloaded text file titled **"reviewChecklist.txt"** for **the second paper** into the text box below.

Survey: Introducing the Empirical Standards Checklist

2. Section 2: Demographic Questions

Please note that Sections 2 to 7 should take you 30 minutes.

* 5. Please select the research methods that are familiar to you.

- [ ] Action research
- [ ] Benchmarking
- [ ] Case study
- [ ] Case survey
- [ ] Data science
- [ ] Engineering research
- [ ] Experiment (with human participants)
- [ ] Grounded theory
- [ ] Meta-science
- [ ] Multimethodology or mixed methods
- [ ] Optimization study (including search-based software engineering)
- [ ] Qualitative survey (i.e. interviews)
- [ ] Quantitative longitudinal study
- [ ] Quantitative simulation
- [ ] Questionnaire survey
- [ ] Repository Mining
- [ ] Systematic literature review
- [ ] None of the above

* 6. How many papers have you published?

- ◯ 0
- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5+

* 7. Have you ever taken a university course that was dedicated to teach you how to perform research?

◯ Yes ◯ No

* 8. Have you ever used a tool or piece of software to review research papers?

◯ Yes ◯ No

If you answered "Yes", please include the names of any other tools or pieces of software you've used to review research papers.

Survey: Introducing the Empirical Standards Checklist

3. Section 3: Understanding the Statements from the Checklist for the first paper.

At this moment, please refer to the open window of the checklist that you completed for the first paper.

* 9. Do you understand all the statements from the **"Essentials" (ES)** section of **the first paper's** checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "ES#: I do not understand the statement because ...".

* 10. Do you understand all the statements from the **"Desirable" (DE)** section of **the first paper's** checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "DE#: I do not understand the statement because ...".

* 11. Do you understand all the statements from the **"Extraordinary" (EX)** section of **the first paper's** checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "EX#: I do not understand the statement because ...".

* 12. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The statements from the checklist for **the first paper** were overall easy to understand.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

4. Section 4: Understanding the Statements from the Checklist for the second paper.

At this moment, please refer to the open window of the checklist that you completed for the second paper.

* 13. Do you understand all the statements from the **"Essentials" (ES)** section of **the second paper's** checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "ES#: I do not understand the statement because ...".

* 14. Do you understand all the statements from the **"Desirable" (DE)** section of **the second paper's** checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "DE#: I do not understand the statement because ...".

* 15. Do you understand all the statements from the **"Extraordinary" (EX)** section of **the second paper's** checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "EX#: I do not understand the statement because ...".

* 16. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The statements from the checklist for **the second paper** were overall easy to understand.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

5. Section 5: Opinions on Usefulness

* 17. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Using the checklists would enable me to review research papers more quickly.	◯	◯	◯	◯	◯
Using the checklists would enable me to review research papers more easily.	◯	◯	◯	◯	◯
Using the checklists would enable me to increase my understanding of research papers.	◯	◯	◯	◯	◯
Using the checklists would enable me to improve how I review research papers.	◯	◯	◯	◯	◯
Overall, the checklists would be useful to review research papers.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

6. Section 6: Opinions on Ease-of-Use

* 18. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall, I find the checklists easy to use.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

7. Section 7: Final Thoughts

* 19. Do the checklists cover all the important aspects needed to review research papers?

◯ Yes ◯ No

If you answered "No", please state why.

20. Do you have any other comments, questions, or concerns about the checklists?

Please click the Done button below if you are done with the survey!

Thank You!

A.3 Survey for Studies Other Than Pilot Study

Survey: Introducing the Empirical Standards Checklist

1. Section 1: Read and Review the 2 Research Papers

Please be aware that the time limit to complete Section 1 is 2 hours.
If you need a reminder on what to do for Section 1, you can watch the video of the demo again by clicking Here.

* 1. The time limit is **1 hour** to review **the first paper.**

- ☐ Download **the first paper** by clicking Here.
- ☐ Read **the first paper.**
- ☐ Review **the first paper** by completing the Empirical Standards Checklist that is listed Here.
- ☐ To complete **section 2** of the survey, it is important to keep the window of the checklist of **the first paper** open.
- ☐ After completing the checklist for **the first paper**, do not forget to download the text file titled **"reviewChecklist.txt"** by clicking the Download button at the bottom of the window.

* 2. Please copy and paste all the text from your downloaded text file titled **"reviewChecklist.txt"** for **the first paper** into the text box below.

* 3. The time limit is **1 hour** to review **the second paper.**

- ☐ Download **the second paper** by clicking Here.
- ☐ Read **the second paper.**
- ☐ Review **the second paper** by completing the Empirical Standards Checklist that is listed Here.
- ☐ To complete **section 3** of the survey, it is important to keep the window of the checklist of **the second paper** open.
- ☐ After completing the checklist for **the second paper**, do not forget to download the text file titled **"reviewChecklist.txt"** by clicking the Download button at the bottom of the window.

* 4. Please copy and paste all the text from your downloaded text file titled **"reviewChecklist.txt"** for **the second paper** into the text box below.

Survey: Introducing the Empirical Standards Checklist

2. Section 2: Understanding the Statements from the Empirical Standards Checklist for the first paper.

At this moment, please refer to the open window of the Empirical Standards Checklist that you completed for the first paper.

* 5. Do you understand all the statements from the **"Essentials" (ES)** section of **the first paper's** Empirical Standards Checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "ES#: I do not understand the statement because ...".

* 6. Do you understand all the statements from the **"Desirable" (DE)** section of **the first paper's** Empirical Standards Checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "DE#: I do not understand the statement because ...".

* 7. Do you understand all the statements from the **"Extraordinary" (EX)** section of **the first paper's** Empirical Standards Checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "EX#: I do not understand the statement because ...".

* 8. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The statements from the Empirical Standards Checklist for **the first paper** were overall easy to understand.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

3. Section 3: Understanding the Statements from the Empirical Standards Checklist for the second paper.

At this moment, please refer to the open window of the Empirical Standards Checklist that you completed for the second paper.

* 9. Do you understand all the statements from the **"Essentials" (ES)** section of **the second paper's** Empirical Standards Checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "ES#: I do not understand the statement because ...".

* 10. Do you understand all the statements from the **"Desirable" (DE)** section of **the second paper's** Empirical Standards Checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "DE#: I do not understand the statement because ...".

* 11. Do you understand all the statements from the **"Extraordinary" (EX)** section of **the second paper's** Empirical Standards Checklist?

◯ Yes ◯ No

If you answered "No", please list the ID with the corresponding reason why you do not understand the checklist statement. For example, "EX#: I do not understand the statement because ...".

* 12. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The statements from the Empirical Standards Checklist for **the second paper** were overall easy to understand.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

4. Section 4: Opinions on Usefulness

* 13. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Using the Empirical Standards Checklists would enable me to review research papers more quickly.	◯	◯	◯	◯	◯
Using the Empirical Standards Checklists would enable me to review research papers more easily.	◯	◯	◯	◯	◯
Using the Empirical Standards Checklists would enable me to increase my understanding of research papers.	◯	◯	◯	◯	◯
Using the Empirical Standards Checklists would enable me to improve how I review research papers.	◯	◯	◯	◯	◯
Overall, the Empirical Standards Checklists would be useful to review research papers.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

5. Section 5: Opinions on Ease-of-Use

* 14. Please say how much you agree or disagree with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall, I find the Empirical Standards Checklists easy to use.	◯	◯	◯	◯	◯

Survey: Introducing the Empirical Standards Checklist

6. Section 6: Demographic Questions

* 15. Please select the research methods that are familiar to you.

- [ ] Action research
- [ ] Benchmarking
- [ ] Case study
- [ ] Case survey
- [ ] Data science
- [ ] Engineering research
- [ ] Experiment (with human participants)
- [ ] Grounded theory
- [ ] Meta-science
- [ ] Multimethodology or mixed methods
- [ ] Optimization study (including search-based software engineering)
- [ ] Qualitative survey (i.e. interviews)
- [ ] Quantitative longitudinal study
- [ ] Quantitative simulation
- [ ] Questionnaire survey
- [ ] Repository Mining
- [ ] Systematic literature review
- [ ] None of the above

* 16. How many papers have you published?

- ◯ 0
- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5+

* 17. Have you ever taken a university course that was dedicated to teach you how to perform research?

◯ Yes ◯ No

* 18. Have you ever used a tool or piece of software to review research papers?

◯ Yes ◯ No

If you answered "Yes", please include the names of any other tools or pieces of software you've used to review research papers.

Survey: Introducing the Empirical Standards Checklist

7. Section 7: Final Thoughts

* 19. Do the checklists cover all the important aspects needed to review research papers?

◯ Yes ◯ No

If you answered "No", please state why.

20. Do you have any other comments, questions, or concerns about the checklists?

Please click the Done button below if you are done with the survey!

Thank You!

A.4 Microsoft Form for Recruitment of Participants

Focus Group to Review the Empirical Standards Checklist

You are invited to be a part of a focus group where you will provide feedback on a checklist, which aims to help Computer Science Master's and Ph.D. students review Empirical research papers.

The event will take about **3 hours** on **Friday, July 22** from **1pm to 4pm (PT)** & you will be rewarded with a **$60 Canadian Amazon Gift Card !**

Please note that we only have a limited capacity for this focus group, so not all of you may be able to attend. We will confirm by email if you have been selected or not to join the focus group. Thank you for your understanding!

* Required

Thank you for your interest in joining our focus group! To complete your registration, please fill out the information below (your personal information will be

1. What is your First Name? *

2. What is your Last Name? *

3. What is your Email Address? *

4. Please state which University you are currently attending. *

5. Are you a Master's student or a Ph.D student? *

- ◯ Master's Student
- ◯ Ph.D. Student

6. Have you published any papers? *

- ◯ Yes
- ◯ No

7. If you have published any papers, please write the titles of the papers below.
   If you have not published any papers, please write None. *

8. What types of Empirical Research have you done? *

- [ ] Action research
- [ ] Benchmarking
- [ ] Case study
- [ ] Case survey
- [ ] Data science
- [ ] Engineering research
- [ ] Experiment (with human participants)
- [ ] Grounded theory
- [ ] Meta-science
- [ ] Multimethodology or mixed methods
- [ ] Optimization study (including search-based software engineering)
- [ ] Qualitative survey (i.e. interviews)
- [ ] Quantitative longitudinal study
- [ ] Quantitative simulation
- [ ] Questionnaire survey
- [ ] Repository Mining
- [ ] Systematic literature review
- [ ] An empirical method not listed above
- [ ] None of the Above

A.5 Research Paper for Study - Qualitative Survey Paper

ABSTRACT

Computational notebooks—such as Jupyter or Colab—combine text and data analysis code. They have become ubiquitous in the world of data science and exploratory data analysis. Since these notebooks present a different programming paradigm than conventional IDE-driven programming, it is plausible that debugging in computational notebooks might also be different. More specifically, since creating notebooks blends domain knowledge, statistical analysis, and programming, the ways in which notebook users find and fix errors in these different forms might be different. In this paper, we present an exploratory, observational study into how notebook users find and understand potential errors in notebooks. We presented users with notebooks pre-populated with common notebook errors—errors rooted in either the statistical data analysis, knowledge of domain concepts, or in the programming. We then analyze the strategies our study participants used to find these errors and determine how successful each strategy is at identifying errors. Our findings indicate that while the notebook programming environment is different from the environments used for traditional programming, debugging strategies remain quite similar. It is our hope that the insights presented in this paper will help both notebook tool designers and educators make changes to improve how data scientists discover errors more easily in the notebooks they write.

1 INTRODUCTION

Jupyter Notebooks[1] are an open-source, browser-based programming environment that allows users to weave together rich text, code, equations, and visualizations into a single human-readable document. Jupyter Notebooks and other computational notebooks, such as Google Colab[2], RMarkdown Notebooks[3], and Azure Notebooks[4], have become immensely popular for anyone who wishes to perform data analysis or exploration tasks. The Jupyter platform specifically has grown exponentially since 2015, with over 6 million publicly available Jupyter Notebooks currently residing on GitHub alone [10].

Despite the popularity of Jupyter Notebooks, its users have mixed opinions about the process of debugging [5]. Some Jupyter Notebook users praise the tool's ability to help them find errors quickly, but others complain about the poor debugging experience [5]. This has given us some insight into how users *feel* about debugging Jupyter Notebooks, but not much about their actual debugging processes.

Debugging involves two phases. The first phase involves identifying what and where the error is, and the second phase involves determining how best to fix the error [**?**]. We have many insights about the process of debugging software systems and programs in general [2, 9, 19], but we lack similar insights for computational notebooks. To the best of our knowledge, there is only one study looking at how computational notebooks are debugged, from Yang *et al.* [**?**], and it focused on a tool for improving code understanding with program synthesis, and did not synthesize strategies used in debugging. Without knowing *how* computational notebooks are debugged, it is difficult to support—whether with tools or processes—the debugging of computational notebooks.

Motivated by the sheer popularity of the Jupyter Notebook platform across many disciplines, we aim to identify strategies adopted by users of Jupyter Notebooks in the first phase of debugging, i.e., strategies for *identifying* errors in data analysis notebooks. Understanding debugging strategies may provide educators and students with valuable knowledge about how errors are found in computational notebooks—improving how they teach or learn error-finding methods. Although our findings are not aimed at any specific community of Jupyter Notebook users, as the platform is used across many different domains [4, 14, 18], we hope that our results may serve as recommendations for users of Jupyter Notebooks who may not have a strong background in the process of debugging.

Our exploratory study is focused on this research question:

What strategies do data scientists use to find statistics, data, and programming errors in Python Jupyter Notebooks?

We performed an observational study with 14 participants who were each tasked with finding and understanding errors in one of four Jupyter Notebooks. We pre-populated errors in the statistics, data (i.e., domain knowledge), and programming code for each of the four notebooks. The participants could use any error-finding method of their choosing and were not time-constrained. Participants we observed followed seven different strategies. We characterized the strategies according to the frequency of use and successfulness (which are not necessarily the same!). Consulting external resources via a search engine was the most *common* strategy. However, the most *successful* strategy was Expectation Confirmation, where there was a mismatch between what explanatory markdown cells claimed and what the code actually did. On average, our participants found approximately 40% of the errors in the notebook they analyzed.

[1] https://jupyter.org/
[2] https://colab.research.google.com/
[3] https://rmarkdown.rstudio.com/
[4] https://notebooks.azure.com/

2 BACKGROUND

Jupyter Notebooks have become so popular that they have been described as the "de-facto standard" for data scientists [11], and much has been learned about how reproducible they are, the quality of the code written in them, and the narratives that describe the analyses within them [12, 13, 16, 17]. These studies agree that notebook code is frequently low-quality and error-prone. Closely related to our work is that from Yang *et al.* [**?**]. They report on a tool to support bug detection in Kaggle notebooks, which they characterized as 'data wrangling code'. Like us, they show this style of development is quite different than pure source code approaches and prone to errors, yet not well supported by tools. They introduce WrangleDoc, a program synthesis technique to summarize code in order to facilitate debugging. We do not examine such summarization approaches in our study, and our Jupyter instance contains no plugins.

While the work of Yang *et al.* highlights the potential problems with data science code, there is still much we do not understand. In particular, their study looked at specific tool support using a documentation approach. We focus instead on the strategies leading to the discovery of notebook errors, the starting point of the debugging process.

Challenges of using notebooks. While the literature on notebook error detection is limited, other research has revealed the challenges of using notebooks. Chattopadhyay *et al.* identified the main pain points of using computational notebooks, including Jupyter [5]. Of interest to this study is the *Manage Code* pain point which mentions that without sufficient software engineering support, debugging, writing code, managing dependencies, and testing relied on *ad-hoc* workarounds. Specifically, they found that writing code in notebooks efficiently requires knowledge of all the function names and classes, plus the use of a second window to search for online resources such as documentation [5]. They also observed a divide in how their participants felt about debugging in notebooks. Some participants were able to find errors in a notebook quickly, but others found that debugging was a horrible experience when they had to rely on `print()` statements. In addition, they found that testing in computational notebooks is difficult as there is no standard method to test a notebook. Some study participants wrote test cases in the same notebook, while others created a new notebook for testing.

Debugging traditional programs. There is plenty of related research on debugging conventional programs. We mention some closely related studies here. Murphy *et al.* presented a qualitative study of the debugging strategies employed by computer science students [9]. They observed three distinct categories of strategy: the good, the bad, and the quirky. The good strategies (or effective strategies) included *gaining domain knowledge, tracing, testing, understanding the code, using resources, using tools, isolating the problem, pattern matching, and considering alternatives.* They also identified that many students employed strategies that were less effective (the bad). These "bad" strategies were the same as the effective strategies, but employed less effectively.

Hypothesis-driven debugging. Alaboudi and Latoza authored two papers that relate to our study. The first paper, titled "Using Hypothesis as a Debugging Aid" [1], describes two studies. In the first study, they observed live stream videos of developers' programming activities. Their second study was a controlled study of 25 participants who were tasked to debug three API misuse problems. Overall, they observed that developers found it challenging to formulate a reasonable hypothesis about a potential error. In the second paper, "An Exploratory Study of Debugging Episodes", Alaboudi and Latoza observed 15 live-streamed programming sessions (in C, C#, Javascript) [2]. They found that developers spent 48% of their programming sessions debugging. They also found that no single activity dominated a debugging session, with developers spending varied amounts of time on different activities. Additionally, they observed significant differences between long and short debugging episodes. Short debugging episodes focused on editing and testing code, while long debugging episodes involved various activities, such as consulting external resources, and inspecting program state in addition to testing and editing.

Student approaches to debugging. Like our study, Whalley *et al.* examined students' thoughts about their debugging process (in non-notebook code). They examined whether reflecting on the debugging process helps students perceive a need for change in their approach, and if they perceive value in a structured, formal debugging process [19]. Whalley *et al.* used semi-structured interviews to answer their research questions. Their analysis uncovered themes about code comprehension, bug location, information gathering strategies, challenges locating bugs, emotions felt during the debugging process, and the value students give to a formal debugging process. When students were asked to reflect on their debugging process, their comments referred to both high- and low-level activities. High-level activities included activities such as reading code, search space reduction, and hypothesis forming. Most of the reflections shared about debugging were about low-level activities, such as where to place print statements, code tracing, and examining function parameters and return values. Many students perceived debugging as inefficient, likely due to the lack of a formal process to follow. One-third of the participants described their debugging process as flawed, and they universally described their hypothesis forming method as imprecise, opting to guess and check instead.

Data science debugging. Debugging traditional programs is well studied and the works discussed above are a very small subset of the work available on debugging. In contrast to the debugging of traditional programs, data science work can be quite different [**?**]. For example, it is a common part of the workflow for scientists to re-run analyses (e.g., as part of exploratory data analysis). This might happen when, for example, removing outliers, or experimenting with different hyper-parameters. Thus, research has looked at supporting such experimental workflows in order to manage versions of notebooks [7**?**], and support cleanup and refactoring [6]. Related to that is work that uses tools to debug data flows in large (non-notebook) data analytics pipelines [**?**], or support transparency in statistical reporting with multiverse analysis [**?**]. None of this work focused on *how* errors were detected, however.

In support of the types of error present in data analyses is the work of Brown *et al.* [3].

3 MATERIALS AND METHODS

To answer our research question, we observed the behaviour of 14 participants as they browsed and debugged existing Jupyter Notebooks that contained errors. The observations took place over Zoom, and participants shared their screens. We recorded video and audio of the meetings for a later qualitative analysis of the strategies our participants used. The study was conducted from November 2020 to January 2021.

3.1 Participants

Table 1 summarizes the participant demographics. A majority (8) of the participants were from the domain of computer science or computer science combined with either music, biochemistry, or life sciences. The other (5) participants were each students in one of chemistry, physics, civil engineering, software engineering, and electrical/computer engineering. The remaining participant was a professional who worked in the education domain. Nine participants had used computational notebooks for less than one year, and the remaining five used them for between one and three years. All participants stated that Jupyter Notebooks was their computational notebook of choice, with two also using Google Colab.

Our participants were recruited by contacting the instructors of three 400/500-level courses with data science themes, by posting on various online communities, and through personal contacts. Participation in our study was voluntary. However, we encouraged participation by providing a $25 Amazon gift card to the first 12 respondents. Our study was approved by our institutional review board.

3.2 Study Materials

The notebooks used for this study were provided to us by a professional educator at RStudio and were initially written in RMarkdown. We translated the notebooks to Python, and the translations were verified by a third party, experienced in Python, and an associate professor of statistics and data science. Additionally, the first author of this paper verified that the Python translations returned the same data, visualizations, and values.

We created two notebooks, each covering a different topic (NBA Player of the Week and the 2011 Spain Election). Each notebook had three versions: A, B, and C. Notebook versions A and B contained errors, while version C did not. Table 2 lists the number of errors per notebook. These notebooks are available in our supplementary material [**?**].

Our notebooks contained the following three types of errors: **data**, **statistical**, and **programming**. These three different types of errors align with the different categories of errors defined by Brown *et al.* [3]. We describe the three categories in the following and provide examples of the different types of errors using the NBA Player of the Week notebook(s).

Data errors occur when the notebook did not fully explore the dataset, or the format of the data was misunderstood.

Listing 1 shows an example of a **data** error: the `Height` column is of type string and is assumed to either contain the centimeter unit or a string value representing a measurement. Similarly, it is assumed that the measurements in the `Weight` column are all in

```
nba = nba.assign(
    Height = pd.to_numeric(nba['Height']␣
    ↪  .str.replace('cm','')␣
    ↪  .str.replace('-[0-9]*',''')),
    Weight = pd.to_numeric(nba['Weight']␣
    ↪  .str.replace('kg','')))
```

Listing 1: A data error: assumes all data is in cm/kg.

```
my_test = ttest_ind(
x1 = nba[(nba['Position'] ==
↪  'PG')]['Height'],
x2 = nba[(nba['Position'] ==
↪  'SG')]['Height'],
alternative = 'smaller')
```

Listing 2: A statistical error: using a 1-sided t-test.

```
nba.groupby('Player').agg(
    Height=('Height', 'median'),
    Weight=('Weight', 'median'),
    Position=('Position', 'first'))
```

Listing 3: A programming error: using the original, not the filtered data-frame.

kilograms, with some measurements containing the kilogram units. In fact, the `Height` column has units of either centimeters, such as 203cm, or feet-inches, such as 6-11. The `Weight` column has measurements which are in kilograms and contain the kilogram units, or measurements in pounds that contain no units. The above code cell removes the centimeter unit by calling `str.replace('cm','')`. The inches measurement is also removed by calling `str.replace␣ ('-[0-9]*','')`. The same is done with the `Weight` column, removing the kilogram units via `str.replace('kg','')`. Thus, the column is incorrectly cleaned as the above code cell performs no unit conversions. This leaves the `Height` and `Weight` columns in mismatched units without any unit identifier.

Statistical errors occur either when an incorrectly chosen statistical test or visualization is used, or when a correctly chosen test or visualization is wrongly interpreted by the user.

In Listing 2, the goal is to determine if a statistical difference between the average height of point guards and shooting guards exists using a t-test. The error is that the `alternative` parameter is set to `smaller`, indicating a one-sided t-test. This parameter should be set to 'two-sided' as the goal was to determine whether or not a statistical difference exists, rather than which average was smaller.

Lastly, **programming** errors occur when a code cell does not achieve the goal stated in the preceding markdown cell. The goal of Listing 3 is to filter the `nba` data-
frame so that it only contains unique players. While this code cell does output a set of unique players, a copy is returned, which is

Table 1: Participant Demographics

Participant	Role	Domain	Notebook Experience (yrs)
P1	Master's	Electrical & Computer Engineering	< 1
P2	Master's	Chemistry	1-3
P3	Undergraduate	Computer Science & Biochemistry	< 1
P4	Undergraduate	Computer Science & Life Sciences	1-3
P5	Master's	Computer Science	< 1
P6	Undergraduate	Physics	1-3
P7	Master's	Civil Engineering	< 1
P8	Undergraduate	Software Engineering	< 1
P9	Undergraduate	Computer Science	1-3
P10	Educational Specialist	Education	< 1
P11	Undergraduate	Computer Science & Music	< 1
P12	Undergraduate	Computer Science & Music	< 1
P13	Doctoral	Computer Science	1-3
P14	Undergraduate	Computer Science & Music	< 1

not saved to the nba dataframe. Through the remainder of this notebook, the original nba dataframe is used, and thus the code has an error and does not achieve its goal.

3.3 Study Task

Each participant was tasked with finding potential errors in either version A or B of the notebook given to them. Version C was shown to them after their analysis if they wanted to see an error-free version. We aimed to balance the number of participants analyzing each notebook (see Table 2). This task was open-ended in that the participants were allowed to use any method they liked to find potential errors. The only specific instructions given were for them to think aloud whenever possible and to notify the researcher when they thought they had found an error.

Table 2: Number of Participants and Errors per Notebook

Notebook	Participants	Errors
nba_analysis_A	4	6
nba_analysis_B	3	4
elections_analysis_A	4	10
election_analysis_B	3	10

3.4 Study Design

We performed two rounds of pilots (with members of our research group) to improve the study task and to confirm our study would provide sufficient observations on error finding strategies. The feedback from the pilots helped us improve the study materials. Our supplementary materials contain the task description, interview questions, and Jupyter Notebooks [**?**].

At the start of each study session, we described the task and emphasized that our aim was not to test their skills. Participants were then presented with a Jupyter Notebook, and they were told that any of the notebook components might contain errors that they should try to identify. We mentioned they may wish to modify the notebook, search the documentation, or use the internet for help.

Each study session consisted of two phases: an observational phase and an interview phase. During the observational phase, the participants analyzed the notebook for errors while one researcher observed their behaviours and took notes. Once the participant was satisfied with their analysis of the notebook, we held the interview phase of the study.

The interview began with unstructured questions, using notes from our observations to guide our follow-up questions. Asking these questions immediately after the participant had performed their task was important as their strategies were still fresh in their mind. These unstructured questions were asked to gain insights into a specific approach and why it was used. Following the unstructured part of the interview, additional questions were asked about the participant's domain of study, how long they had been using Jupyter Notebooks, and the computational notebook they used most often. The complete list of these additional questions is available in our supplemental package (as mentioned above).

3.5 Data Collection and Analysis

The Zoom video recordings of the studies were uploaded, and we analyzed the recordings directly using ATLAS.ti 8[5]. We used an open coding process to code all activities performed by our participants. The first author of this paper performed the initial coding. After the initial coding cycle, discussion sessions were held with the second and fourth authors, where the codes were further analyzed and compared with the findings from previous participants, and refined in an iterative manner.

Throughout our discussion sessions, we identified emergent higher-level groups for the codes and merged some codes where necessary, shown in the list below:

- **Action:** An action that a participant performed.
- **Docs:** A specific documentation website that a participant visited.

[5]https://atlasti.com/product/what-is-atlas-ti/.

- **Online Resource:** An online resource other than documentation that a participant visited.
- **Reasoning:** A reason for performing an action or a reason for why something was an error.
- **Participant Attribute:** An attribute that describes a participant.

Throughout our discussion sessions, we noticed many *actions* were performed together. We called these connected sets of actions *strategies*. The first author analyzed the raw data again to identify and code strategies from each group of *actions*.

Once strategies had been identified, the videos were analyzed again in order to determine the success rate of each strategy. Whenever a participant was analyzing an erroneous cell, their chosen strategy was entered into a spreadsheet, along with the type of error they were working on, and if that strategy was successful or not.

4 FINDINGS

Our research question asked what strategies data scientists use to find statistics, data/domain, and programming errors. From our analysis, we identify: (a) a set of *actions* and (b) *strategies* that our participants employ to find errors in Python Jupyter notebooks. Additionally, we present (c) the *relationship between strategies and error-finding success*. Tables 3 and 5 show the entire list of actions and strategies identified in our exploratory observational study. Finally, in Table 6, we present how strategies are related to the different error types.

We describe each action, strategy, and their respective relationship with an error type in the following.

4.1 Actions Taken

Our participants performed various actions while analyzing the notebooks to find errors. In this context, an action is an (atomic) activity such as reading a markdown/code cell or examining a CSV file. Table 3 lists the number of participants who performed each action and the average amount of time all participants spent per action. There were some actions that participants always used, such as reading code and markdown cells, writing or editing code, and using the search engine. Other actions often used were looking at the documentation, checking code output, and inspecting dataframes. Finally, a few actions were only occasionally used, such as inspecting a CSV file or adding a comment. We describe these actions in more detail below.

A1: Reading a code cell. This action refers to when participants (P1-P14) *read through a code cell to understand what it was doing.*

A2: Reading markdown cells. This action refers to when a participant (P1-P14) *read through a markdown cell to gain context into what the preceding code cell tried to accomplish.*

When analyzing a notebook to find errors, participants performed actions A1 and A2 successively. In this scenario, P8 emphasizes *"[I read] the documentation first then [I read] the code"*. Additionally, P12 highlighted the value of reading the markdown aloud to better understand what was going on.

A3: Writing/Editing code. This action occurred when participants (P1-P14) *wrote new code in a code cell (either one they added or one present in the notebook) or edited a code cell that was initially in the notebook.* We observed that participants edited code for several different reasons. For example, P14 stated that they edited code cells to make them more readable. Other participants, such as P10, edit the parameters of functions to view more of the data returned by that function; for example, P10 edited calls to Pandas `Series.nlargest()` function. Some participants also wrote new code into the notebooks, which served various purposes. For instance, P13 wrote code during their analysis of the nba_analysis_A notebook to verify if two sets of rows in the `nba` dataframe were the same. Both P6 and P11 wrote code to perform type checking through the use of Python's `type()` method.

A4: Using a search engine. All participants used a search engine to *access some online resource or documentation page.* Typically participants transitioned from the notebook to the search engine and then to either an online resource or a documentation page. Depending on whether or not the initial search result was helpful, they would return to the notebook or select another result from the search engine. The **Search Engine** action is highly associated with both **A5: Looking at documentation** and **A8: Looking at an online resource**. We define online resources as any website other than a documentation page. Table 4 shows the most commonly accessed documentation websites and online resources.

A6: Checking code output. Commonly, participants (P1-P12, P14) *inspected the output of a code cell visually, either one initially present in the notebook or one which the participant added.*

A7: Inspecting dataframe. We observed that participants (P1-P4, P6-P12, P14) used the `Dataframe.head()` method to *visually inspect the dataframe, either to gain a preliminary understanding of the data or to check if anything seemed out of place.*

A9: Inspecting a graph. Participants (P3, P5-P14) performed this action to *visually inspect any graph present in the notebook.*

A10: Inspecting CSV File. In a similar situation to A7, participants (P5, P6, P9, P10, P13) inspected the data in its raw state. For instance, P13 pointed out that when using Jupyter Notebooks, they do not use the CSV viewer native to Jupyter; instead, they use an alternative application. Likewise, P9 indicated they use Notepad++ to view their CSV files. Finally, P10 highlights that they inspected the CSV file when they were unsure how to perform a task programmatically. In this matter, P10 states *"I'm just learning Python, so I can't...list these things, I actually refer to the CSV quite a bit"*.

A11: Reading an error message. This action occurred when participants (P1, P3, P5, P8-P11) changed the notebook as initially, the notebooks did not return any error messages. In this scenario, P10 points out that when they see an error message, they *"don't have a clue"*.

A12: Adding a comment. This action occurred when participants (P5, P11, P14) *added a comment to a code cell either in the form of a note or to comment out code.*

While these actions capture the more atomic tasks our participants performed, we also observed that several actions were used together to form strategies that helped participants find or understand the cause of errors. In the remainder of this section, we describe these strategies in more detail.

Table 3: Actions Taken in Error Identification.

Action	Action ID	# Participants	Average Time Spent (mm:ss)
Reading code cell	A1	14	06:46
Reading markdown	A2	14	05:51
Writing/Editing code	A3	14	03:21
Using a search engine	A4	14	01:55
Looking at documentation	A5	13	03:00
Checks code output	A6	13	02:24
Inspecting DataFrame	A7	12	04:07
Looking at an online resource	A8	12	03:06
Inspecting graph	A9	11	01:53
Inspecting CSV file	A10	5	02:40
Reading an error message	A11	3	01:51
Adding a comment	A12	2	08:42

Table 4: Number of Visits. † indicates a Documentation page. The remainder are Online Resources. Fourteen other Online Resources were each visited between one to three times.

Resource	Number of Visits
Pandas †	58
Plotnine †	28
Statsmodels †	21
stackoverflow.com	14
geeksforgeeks.org	6
investopedia.com	6
Numpy †	4
Scipy.stats †	4
tutorialspoint.com	4
w3schools.com	4

4.2 Error Finding Strategies

Participants performed many of the preceding actions together to serve a particular purpose. We call a collection of related actions a *strategy*. We describe the strategies we found in detail. Table 5 gives a brief description along with the number of participants who used this strategy.

Search Engine Driven Approach. The most common strategy we observed was the *search engine-driven approach*, which every participant used. All participants made several transitions from the notebook to the search engine, then to an external resource, until they found a helpful online resource or documentation page. Participants outlined three different reasons for using the search engine and external resources: First, using the search engine as a first step to gather a solution from an online reference. For instance, P12 highlights that they use Google quite often when using a Jupyter Notebook, and without it, they would not know what to do. Not knowing what to do without the search engine hints at being dependent on it; it is unknown whether this is caused by a lack of general programming knowledge or knowledge of a specific API, such as Pandas.

Second, the search engine was also used as a confirmatory aid; this happened when participants had prior knowledge. However, they seek supplementary expertise to confirm or refresh their intuition. For instance, P7 stated they often remember general concepts but use the search engine to gather information about what some specific terms mean to interpret them correctly, such as when P7 gathered information about interpreting the results of an ordinary least squares (OLS) regression.

Finally, participants may use the search engine to gather code snippets as potential solutions. P8 emphasized that their particular use of the search engine was to find code snippets that help them fix the errors they have identified.

Assume and (Sometimes) Check. Participants would only cursorily inspect a code cell, see what the code is doing, and return to it only when they identified a potential problem in their theory of the notebook's execution. They then made an assumption about where in the preceding cells that problem happened. They then examined that code in more detail than they did on their first pass over it. However, participants "sometimes" left some assumptions unchecked. This may be due to the contrived nature of the study (fixing the bug was not part of the task). When participants did check assumptions, they wrote new code in the notebook or examined the dataframe/CSV file.

Consider the error and thought processes of P8 while they use the *assume and (sometimes) check* strategy to determine the error described in Listing 1 (code cell 3 of the notebook NBA_Analysis_A). P8 began analyzing the notebook using a *once-over* (see the next strategy) and notices in a later code cell that the given mean of the height column is roughly 12. They then remark that a *"mean height of 12 doesn't seem to make a lot of sense"* (since height in cm should be (broadly) greater than 100cm and less than 225cm). They then transitioned to read code cell 3 (Listing 1 line 2), which cleaned and adjusted the height column. Rereading the code cell led them to *assume* something must have gone wrong in that notebook cell. They then inspected the original dataframe and made another assumption: *"Here the measurements are presumably in feet-inches and over here we have them in cm"*. This second assumption is an example of assuming the purpose of a series of method calls. A closer

Table 5: Strategy Descriptions

Strategy	Description	# Participants	Associated Actions
Search Engine Driven Approach	Using the search engine and external resources to gather useful information.	14	A4, A5, A8
Assume and (Sometimes) Check	Making an assumption related to the notebook or to an API call and sometimes checking it.	14	A3, A7, A12
Expectation Confirmation	The participant's expectation, set up by an explanatory markdown cell, of what a code cell does cannot be confirmed upon seeing its output.	7	A1, A2
Once-Over	Briefly browsing through the notebook in order to gain a preliminary understanding of what it contains.	4	A1, A2, A6
Re-implement to Check	Re-implementing a code cell using a different syntax in order to check its validity.	3	A3, A6
Key Information	Extracting need-to-know information from a markdown cell and placing it in a comment inside the related code cell.	1	A2, A10
Start With What You Know	Starting at a point in the notebook which is most familiar.	1	A1, A2

inspection of code cell 3 allows them to identify the error as replacing inches with the empty string and not accurately converting feet-inches to cm.

Expectation Confirmation. Seven participants (P1, P3, P5, P7, P10, P11, P13) indicated a discrepancy between explanatory text in a markdown cell and the subsequent code cell helped them identify an error. P7 described the explanatory markdown as a *"guidance for what I should be looking for"*, and that when a difference occurs between the markdown and the code cell, they know something is incorrect. Additionally, P5 used an analogy to describe the discrepancy between the markdown and code, stating, *"It's basically like 'Hey, we did this' and then [I] look at the code and it's like 'No, you didn't.'"* Finally, P11 emphasized, *"what I was expecting is that we want a percentage and this is obviously not a percentage"*, outlining how the markdown sets their expectations. When the code does not fulfill these expectations, they know something is wrong.

Once-Over. Four participants (P2, P6, P8, P14) used this strategy, which involves looking through the notebook to gain a preliminary understanding. This strategy consists of reading markdown and code cells, running code cells and briefly checking their output, and generally inspecting the notebooks' initial state. A once-over gives a basic understanding of what the notebook is doing without too much detail. All four of the participants, when using the *once-over* strategy, employed different language to describe it. For example, P2 stated they were getting *"a lay of the land"*.

Re-implement to Check. The *re-implement to check* strategy was used by three participants (P1, P6, P11) and implies rewriting a code cell using a different syntax and then comparing the results of both to see if there are any differences. For example, P1 wrongly believed that Listing 4 was incorrect due to the `.agg()` syntax.

They went on to add a new cell and rewrite the code (Listing 5), only to find that they produced the same result.

P6 stated that they would have shown a correlation by plotting rather than using an OLS regression. However, they did not

```
nba[(nba['Position'] == 'PG') |
    (nba['Position'] ==
    'SG')].groupby('Position')
    .agg(Height=('Height', 'mean'))
```

Listing 4: Code snippet P1 wrongly thought was incorrect.

```
nba[(nba['Position'] == 'PG') |
    (nba['Position'] ==
    'SG')].groupby('Position')
    .agg('Height').mean()
```

Listing 5: P1's re-implementation of Listing 4.

re-implement this code cell as they were unfamiliar with the plotnine package used to generate the plots. While not precisely re-implementation, P11 would write pseudocode before looking at a code cell and after reading its markdown explanation. They would then compare this pseudocode to the actual code, and if they were similar, P11 believed this code cell was correct and moved on to a new cell. Additionally, participants could combine this pseudocode strategy with an actual re-implementation to further validate a given code snippet.

Key Information. The *Key Information* strategy was used four times by P5 and describes extracting only the information you need from the markdown description of a code cell; P5 then placed this information inside the code cell as a comment. Extraction of the key information allowed P5 to get the information closer to the code, and reduced the number of times they re-read a markdown cell to remind themselves of what a code cell is doing. In addition, they highlighted how extracting the key information allowed for easier comparison of the code and markdown, and eliminated any extraneous information they did not need to know. Using the

Table 6: Relating Strategies (from Table 5) and Error Type. A dash (-) indicates no use. The once-over strategy was not used for any of the error types. There are a maximum of 21 programming errors, 12 statistical errors, and 7 data errors.

Strategy	Error Type	Times Used	Errors Found	Percentage
Search Engine Driven Approach	Programming	**26**	11	52.4%
	Statistical	**16**	**7**	53.9%
	Data	**10**	2	28.6%
Assume and Check	Programming	13	7	33.3%
	Statistical	8	4	30.8%
	Data	5	4	57.1%
Expectation Confirmation	Programming	20	**17**	81.0%
	Statistical	6	0	0%
	Data	7	**7**	100%
Re-implement to Check	Programming	1	0	0%
	Statistical	-	-	-
	Data	1	0	0%
Start With What You Know	Programming	1	0	0%
	Statistical	-	-	-%
	Data	1	0	0%
Key Information	Programming	3	2	9.52%
	Statistical	2	1	7.69%
	Data	-	-	-%

Key Information strategy allowed P5 to more easily employ the *Expectation Confirmation* strategy.

Start With What You Know. P5 employed another strategy named *Start With What You Know*, which involved analyzing parts of the notebook they were familiar with first. They mentioned that doing so made them *"feel more confident"*, and that starting with the topics they are more familiar with gave them a better chance to find errors. This confidence then allowed them to find errors in the other sections of the notebook as they were better able to understand the nature of the errors.

We now describe how the strategies outlined in Section 4.2 were used to find the various types of errors present in each notebook. As our study is exploratory, we do not make any claim that these are the best strategies for finding a particular type of error (such a claim would require future work). Recall that our study included the analysis of three types of errors: programming errors, statistical errors, and data/context errors (see Section 3.2). The error types are not mutually exclusive, and a given error can belong to more than one error type. Table 6 outlines the number of times our participants used each strategy per error type, the number of errors found per strategy, and the percentage of total errors found by each strategy. We report on all seven strategies. The most successful strategies are **Expectation Confirmation** and **Search Engine-Driven Approach**. The **Expectation Confirmation** strategy success is influenced by the markdown present in our notebooks. The markdown description sets up expectations for our participants. When the participants read the code following the descriptive text, they contrast their expectations of what the code is supposed to do with what the code actually does. We note that in practice, Pimentel *et al.* found that notebooks contain very little markdown [12].

Additionally, we note that the efficacy of the **Search Engine-Driven Approach** is associated with the popularity of using online resources to guide users of Jupyter Notebooks [8], as pointed out by participants P7, P8, and P12. Koenzen *et al.* similarly determined that code reuse in Jupyter Notebooks most commonly came from searches on the web, most often from websites that provided a tutorial, followed by API documentation [8].

5 DISCUSSION

This section discusses the implications of our work to Jupyter notebook users, notebook tool designers, and educators. We also provide insights about differences in debugging notebooks and non-notebook code and the threats to the validity of our work.

5.1 Implications

The error-finding strategies we have identified point to the need for more tool support when developing Jupyter Notebooks, or rather more awareness about the support offered by the tool. This need for more tool support is suggested by other studies as well [5**?**]. For example, all of our participants chose to use a search engine to access external resources, with thirteen using the search engine to read API documentation and online resources. The use of the search engine, however, required context switching to a different browser tab. Note, if our participants used Jupyter Lab, switching to a different browser tab would not have been needed, as one can access Python's docstring class from within Jupyter Lab by using the Shift + Tab shortcut. Similarly, one can also use the Tab shortcut to open the auto-complete functionality of Jupyter Lab.

The release of Jupyter Lab 3.0 introduced a visual debugger that can be used to step through code, or to check the value of a

variable [15]. Other tool support, such as linting, does not come packaged with the latest version of Jupyter Lab but is still available through extensions such as flake8 [6] or the program synthesis approach from Yang *et al.* [**?**].

Knowledge of the support available within Jupyter Lab or through extensions could be leveraged by educators using computational notebooks as part of their curriculum. For instance, if an educator were to make these features of Jupyter Lab apparent to their students, it may help them focus on the task at hand rather than spending time finding the correct version of a documentation page. Additionally, knowing about the strategies their students may use, possibly the ones identified in this paper, may allow them to better guide their students through the process of debugging a computational notebook. Furthermore, notebook tool designers could use design tools similar to Seahawk [**?**] which could integrate online resources like StackOverflow directly into the notebook environment.

5.2 Comparing Debugging Notebooks and Debugging Non-Notebook Code

The development of non-notebook code differs from the development of computational notebooks. The type of problem managed in a notebook involves more data wrangling, experimentation, and analysis code. Our study separated these into potential problems with statistics, programming, and data / domain knowledge. The notebook environment also has a literate programming component that goes beyond code comments, with markdown cells that can be used to describe the purpose of the code.

Furthermore, non-notebook IDEs have robust tool support for debugging, for example, setting breakpoints in IntelliJ. However, in the traditional computational notebook interface (say Jupyter Notebook), debugging is not specifically supported by the tool. Data science tools are actively working to fix this, for example Jupyter Lab's debugger [15] and RStudio's debugging interface. IDEs are also now able to integrate notebook code into the IDE directly, such as with Visual Studio Code.

Given these differences, we ask whether error identification approaches are also different. This study identified several strategies participants used to identify errors in Jupyter Notebooks. Other researchers have identified strategies for debugging non-notebook code. Table 7 outlines strategies identified by Murphy *et al.* and Whalley *et al.*, that are similar to those we have identified [9, 19].

The **Search Engine-Driven Approach** strategy is closely related to the **Using Resources** strategy identified by Murphy *et al.* in [9]. Both strategies involve the use of documentation and tutorials. The difference between these two strategies is that we observed our participants using the search engine as their gateway to many resources. Murphy *et al.* make no mention of the search engine.

Both the **Information Gathering** and **Bug Location** strategies identified by Whalley *et al.* mention the use of speculation and guessing about the locations and causes of bugs. Alaboudi and LaToza also report on using hypotheses as a debugging aid [1]. Our **Assume and (Sometimes) Check** strategy is similar, based on making an assumption and optionally checking the assumption.

[6] https://github.com/mlshapiro/jupyterlab-flake8

Table 7: Strategies Similar to Those We Identified

Strategies We Identified	Similar Strategies
Search Engine Driven Approach	Using Resources [9]
Assume and (Sometimes) Check	Information Gathering [19], Bug Location [19]
Expectation Confirmation	Pattern Matching [9]
Once-over	Gain Domain Knowledge [9], Understanding the Code [9], Static Code Comprehension [19]
Re-implement to Check	N/A
Key Information	Understanding Code [9], Static Code Comprehension [19]
Start With What You Know	N/A

In their description of the **Pattern Matching** strategy, Murphy *et al.* state that their participants found bugs due to things not "looking right". In our notebook study, this was made more explicit than the heuristics Murphy *et al.* describe. Our participants were able to identify errors when a code cell seemed like it was not correct based on a description given in a markdown cell (the **Expectation Confirmation** strategy). The breakdown of an expectation could be thought of as pattern matching as our participants were attempting to match the pattern of what they were told the code was trying to accomplish to what they could observe the code doing. However, the presence of explicit documentation makes this strategy quite successful (at least for our example notebooks).

The **Once-over** and **Key Information** strategies identified by us are both similar to the **Understanding Code** and **Static Code Comprehension** strategies identified by Murphy *et al.* and Whalley *et al.* respectively. In addition, the **Once-over** strategy is similar to the **Gain Domain Knowledge** strategy identified by Murphy *et al.* Both of our strategies were used in order to gain understanding about the contents of the notebook, and involve comprehending both the code and the markdown, much like the **Understanding Code** and **Static Code Comprehension** strategies are about comprehending code. The **Once-over** strategy was used to gain domain knowledge in the sense that the four participants who used this strategy did so to gain a brief understanding of the domain the notebook covered.

We found no similar strategies to the **Re-implement to Check** strategy (modifying the code to verify the output) or the **Start With What You Know** strategy we observed. One advantage of notebooks is that code cells are capable of independent output, closer to a REPL session than debugging a complete source file. This makes these strategies more viable as the code is more granular and independent.

We found that debugging non-notebook code differs from debugging computational notebooks in a few ways. One, the type of development is different: there are more data science related tasks such as data wrangling. Two, the development tools themselves are at different levels of maturity when it comes to debugging support. Three, while five out of seven of the strategies we observed

are related to non-notebook code debugging strategies identified in the literature, we found that two strategies were not found in non-notebook code studies. We also saw differences in how **Expectation Confirmation** and **Assume and (Sometimes) Check** are conducted in practice, given the way a notebook isolates individual code cells.

5.3 Threats to Validity

In the following, we address the validity of this study in the context of qualitative research [**? ?**].

Internal validity We did not impose time-constraints on our participants, and they were assured our study was not a test of their skill. However, given the nature of the task, it is possible our participants felt pressure to perform well. Due to this pressure, participants may have overlooked errors in the Jupyter Notebooks. However, during the interviews we conducted immediately after the tasks, we did not detect that our participants felt any undue stress due to the study.

Construct validity Our study prompt and task description may have influenced participants to perform actions which were not part of their typical error identification process in Jupyter Notebooks. For instance, modifying the notebook, searching documentation, and using the internet for help may not have been naturalistic behaviours. To mitigate this threat, we adopted multiple strategies, such as two rounds of pilots, to ensure the comprehensibility and raise the realism of the tasks. In addition, task descriptions and scripts were reviewed and validated by a domain expert and the task was confirmed to be within the recruited participants' skill level.

External validity The primary threat to external validity is how we recruited and selected participants. We used convenience sampling methods to recruit participants from upper-level undergraduate and graduate level courses at the university. Therefore, most of our participants were students who used Jupyter Notebooks for school assignments and not professionally. However, we designed the tasks according to our participants' skill levels and the context of tasks is fairly approachable (elections and sports) by any participant independently of their academic background. Our participants did not express unfamiliarity with the domain. However, some participants expressed unfamiliarity with specific packages imported into the notebook, namely Pandas and Plotnine.

Reliability The open coding process was performed by one researcher, the first author of this paper. To reduce potential researcher bias and subjectivity, we conducted several discussion sessions to iteratively build a codebook. We confirmed with the feedback of an expert reviewer, the fourth author of this paper, to raise the reliability and maturity of our findings.

6 CONCLUSION

We conducted an observational study with fourteen participants, mostly university students from varying technical backgrounds, and observed the strategies these Jupyter Notebook users employed to identify errors that were seeded in two sample notebooks. Our participants performed a variety of actions while studying the notebooks and looking for errors. When several participants performed a set of these actions in succession, or one participant performed a this set of actions multiple times, we identified the sequence of actions as a particular debugging strategy.

The most commonly used strategy we observed was using the search engine to find external help such as API documentation or websites that gave a tutorial. However, the most successful strategy was Expectation Confirmation, when they discovered a mismatch between the description and the code itself. We identified some implications for practice, including the need for better debugging support in notebooks, and showed that while there are similarities with non-notebook code, debugging in notebooks leverages notebook-only properties such as code cell independence and hidden state. We hope our study design and insights will help both notebook tool designers and educators make changes to improve how data scientists discover errors more easily in the notebooks they write.

ACKNOWLEDGMENTS

Omitted for anonymity.

REFERENCES

[1] Abdulaziz Alaboudi and Thomas D. LaToza. 2020. Using Hypotheses as a Debugging Aid. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, Dunedin, New Zealand, 1–9. https://doi.org/10.1109/VL/HCC50065.2020.9127273

[2] Abdulaziz Alaboudi and Thomas D. LaToza. 2021. An Exploratory Study of Debugging Episodes. *CoRR* abs/2105.02162 (2021). arXiv:2105.02162 https://arxiv.org/abs/2105.02162

[3] Andrew W Brown, Kathryn A Kaiser, and David B Allison. 2018. Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2563–2570.

[4] Alberto Cardoso, Joaquim Leitão, and César Teixeira. 2019. Using the Jupyter Notebook as a Tool to Support the Teaching and Learning Processes in Engineering Courses. In *The Challenges of the Digital Transformation in Education*, Michael E. Auer and Thrasyvoulos Tsiatsos (Eds.). Springer International Publishing, Cham, 227–236.

[5] Souti Chattopadhyay, Ishita Prasad, Austin Z. Henley, Anita Sarma, and Titus Barik. 2020. *What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376729

[6] Andrew Head, Fred Hohman, Titus Barik, Steven M Drucker, and Robert DeLine. 2019. Managing messes in computational notebooks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[7] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The Story in the Notebook: Exploratory Data Science Using a Literate Programming Tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3173574.3173748

[8] Andreas P. Koenzen, Neil A. Ernst, and Margaret-Anne D. Storey. 2020. Code Duplication and Reuse in Jupyter Notebooks. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, Dunedin, New Zealand, 1–9. https://doi.org/10.1109/VL/HCC50065.2020.9127202

[9] Laurie Murphy, Gary Lewandowski, Renée McCauley, Beth Simon, Lynda Thomas, and Carol Zander. 2008. Debugging: the good, the bad, and the quirky–a qualitative analysis of novices' strategies. *ACM SIGCSE Bulletin* 40, 1 (2008), 163–167.

[10] Peter Parente. 2014. Estimate of public jupyter notebooks on github. https://github.com/parente/nbestimate

[11] Jeffrey M Perkel. 2018. Why Jupyter is data scientists' computational notebook of choice. *Nature* 563, 7732 (2018), 145–147.

[12] João Felipe Pimentel, Leonardo Murta, Vanessa Braganholo, and Juliana Freire. 2019. A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE, Montreal, QC, Canada, 507–517. https://doi.org/10.1109/MSR.2019.00077

[13] Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. *Exploration and Explanation in Computational Notebooks*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173606

[14] Adam A Smith. 2016. Teaching computer science to biologists and chemists, using jupyter notebooks: tutorial presentation. *Journal of Computing Sciences in Colleges* 32, 1 (2016), 126–128.

[15] Jeremy Tuloup. 2021. JupyterLab 3.0 is released!. The 3.0 release of JupyterLab brings… | by Jeremy Tuloup | Jupyter Blog. https://blog.jupyter.org/jupyterlab-3-0-is-out-4f58385e25bb (Accessed on 07/28/2021).

[16] Jiawei Wang, Tzu-yang Kuo, Li Li, and Andreas Zeller. 2020. *Restoring Reproducibility of Jupyter Notebooks*. Association for Computing Machinery, New York, NY, USA, 288–289. https://doi.org/10.1145/3377812.3390803

[17] Jiawei Wang, Li Li, and Andreas Zeller. 2020. Better Code, Better Sharing: On the Need of Analyzing Jupyter Notebooks. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results* (Seoul, South Korea) *(ICSE-NIER '20)*. Association for Computing Machinery, New York, NY, USA, 53–56. https://doi.org/10.1145/3377816.3381724

[18] Charles J Weiss. 2020. A Creative Commons Textbook for Teaching Scientific Computing to Chemistry Students with Python and Jupyter Notebooks. *Journal of Chemical Education* 98, 2 (2020), 489–494.

[19] Jacqueline Whalley, Amber Settle, and Andrew Luxton-Reilly. 2021. *Novice Reflections on Debugging*. Association for Computing Machinery, New York, NY, USA, 73–79. https://doi.org/10.1145/3408877.3432374

A.6 Research Paper for Study - Repository Mining Paper

ABSTRACT

Stack Overflow is a rich source of questions and answers—discussions—about software development. One topic of discussion is software design, such as the correct use of design patterns, or best practices in data access. Since design is a more abstract topic in software engineering, researchers have long sought to characterize and model design knowledge. However, these approaches typically require significant expert input in order to contextualize the abstract design information. In this study, we explore how combining expert input with Stack Overflow might serve as an effective way to identify design topics. We first perform a qualitative analysis of design-tagged Stack Overflow questions and answers to identify the design concepts developers discuss. We report on areas where agreement was a challenge, including abstraction levels. Since inductive coding is expensive, we apply a semi-supervised (Anchored CorEx) approach. We find it performs as well as LDA but offers superior interpretability and the ability to guide the topic model. We leverage CorEx to characterize how design is discussed in Stack Overflow and on GitHub. We conclude by describing how our experience using the semi-supervised CorEx approach leads us to believe that approaches like CorEx that combine domain knowledge and scalability are key for analyzing large SE text repositories.

1 INTRODUCTION

Concerns about software design remain difficult to understand. This is because design is one of the least concrete elements in a software lifecycle, at least compared to testing, implementation, and deployment [6]. Software design is usually done heuristically, by pulling from the current project context (e.g., the architecturally significant requirements), the design team's knowledge, and an ever-changing set of patterns and styles from the literature. How this process works, what makes one design good and another bad, and how to support developers in design tasks is subjective and frequently changes as new technologies emerge. In order to examine how developers discuss design topics, we embarked on an exploratory study using Stack Overflow[1] (SO).

We begin by assessing the **best way to extract and characterize design-related knowledge** in Stack Overflow. We first create a dataset of design-related posts. We then compare and contrast three methods of classifying Stack Overflow posts: (1) qualitative, manual, inductive coding; (2) using existing Stack Overflow user tags; and (3) unsupervised and (4) semi-supervised topic modeling. Each of these four approaches can potentially classify Stack Overflow posts into design related topics. We explore the differences in design topics each method identifies. We show how using a semi-supervised approach can bring the best of both highly scalable LDA approaches and the domain knowledge of inductive coding.

We then report on what **design topics are being discussed on Stack Overflow**. We propose some ways in which knowing these topics can be used in a practical way. One is to use these topics to identify what (design-related) questions are being linked in developer source code on GitHub. If developers are already using the Stack Overflow links in repositories for documentation, then this implies tools and processes could be enhanced to make this process even more effective.

The biggest challenge to characterizing latent topics on Stack Overflow is that software design discussions, decisions, and knowledge encompass a broad scope of topics, including sub-areas such as architecture choices, design patterns, and data access, to name a few. Design knowledge can be specific to a particular API or language (e.g., JavaScript's Promises framework), or generic for many

[1] https://stackoverflow.com/

possible implementations (such as Publish-Subscribe). Recovering design knowledge involves finding design information which is often scattered and poorly organized in software projects [10].

However, the emergence of Stack Overflow as a knowledge repository for many software-related topics, such as energy management [28], suggests that software design knowledge also occurs there. Indeed, Soliman et al. [34] show that this is true for the specific case of architecture knowledge for middleware. We expand on their study for the more general case of software design knowledge and derive various software design-related areas that the developers find challenging. For this, we extend the Stack Overflow design-related dataset, created by a previous study [25], using a *Tag Filtering* approach. Using the question tags from SO allows us to leverage a low-cost annotation directly created by the developers themselves [36]. We conduct an exploratory study with the following research objectives:

RQ1: What are effective techniques for classifying design related posts on Stack Overflow? We categorize design-related posts using four approaches. For **RQ1.1** we examine posts using a manual inductive coding approach. We explore how accurate this approach is, and find that it can be very challenging to get agreement. In **RQ1.2** we use an unsupervised approach, LDA topic modelling [9] and an anchored, semi-supervised approach called Anchored CorEx [15, 31, 38]. We compare these two approaches in their precision and recall, and contrast them with the inductive labels we created earlier, and user tags from Stack Overflow.

RQ2: How can we leverage knowledge of design topics discussed on Stack Overflow? While there are many applications of the semi-supervised CorEx approach, we look at two practical implications of the results of RQ1: we apply the semi-supervised design classification approach arrived at in RQ1 on the entire Stack Overflow dataset. We discuss the topic frequency and co-occurrence. Since developers can refer to Stack Overflow in software artifacts such as code comments, we investigate to what extent such links are to design discussions as identified above, and characterize the type of discussions referred to.

We conclude by discussing some challenges in using a semi-supervised topic modeling approach, and limitations of our work.

This paper contributes the following: One, we demonstrate a semi-supervised learning approach to extracting design topics from Stack Overflow, release our tagged dataset from Stack Overflow, and our inductively coded set of anchor terms. Two, we explore practical applications of the resulting topic model, including a characterization of the design discussions on Stack Overflow, including co-occurring topics, and a way to use GitHub to reveal implementation links to design topics. Finally, we compare different topic modeling approaches to prepare guidelines for when and how to apply CorEx.

2 BACKGROUND AND RELATED WORK

Related work for this paper is from research on mining Stack Overflow and related sites, and research about extracting and managing design knowledge.

Mining Stack Overflow. Stack Overflow is the most popular question and answer (Q&A) forum online [3]. It has seen many papers and mining challenges (including the MSR 2015 challenge) from researchers studying it, ranging from software security [42], machine learning [4], energy consumption [28], or non-functional requirements [43], among others. These studies adopt a similar approach to ours in using text analysis tools such as LDA [9] to analyze the dataset, once it is filtered using keyword labels.

Two studies look at how GitHub artifacts and Stack Overflow content are linked. Yang et al. [41] looked at how Stack Overflow code snippets (specifically Python) appeared in GitHub code repositories in a large-scale study. They found evidence that 1-2% of code from Stack Overflow appears in GitHub, which suggests transfer is occurring via copy and paste programming; this implies a similar mechanism may occur for the design concepts we detect (but which would be more challenging to uniquely identify in code). Specific links from code to Stack Overflow are easy to identify, by contrast, and are explicitly identified in the SOTorrent dataset. Manes and Baysal [26] show this occurs quite often, with a median of 144 references per project. Design was not one of the more popular topics they identified, but like us they found developers seem to point more often to the question as way of reference for code changes.

Stack Overflow is the largest development Q&A forum, but has limitations: it may only be used by a subset of practitioners, is English-language, and omits questions that are proprietary or internal (e.g., referring to internal APIs). As user-generated content it is subject to the fallibilities of its users, including off-topic comments, cliques, and bias toward majority groups [13]. Other developer forums are Gitter [12] and GitHub Discussions [19], which may be relevant in future design mining studies.

Finding and managing design knowledge. In the context of software engineering, design is traditionally understood both as a process [14] in which a development team engages, and as the resulting specifications [29] that the team produces.

The idea that useful design information may occur in discussions is supported by the literature on design rationale [24].

Organizing design information was a primary motivation behind the patterns community. Until quite recently, this knowledge was organized largely manually, both because the information was heavily context-specific, but also due to a lack of large datasets. More recently, Gorton et al. semi-automatically populated design knowledge from internet sources for a particular (big data) domain [18], Tian et al. [37] used Stack Overflow to find how architecture smells were discussed, Bi et al. [8] looked at architecture patterns mentioned on Stack Overflow, and Ali et al. [2] uses Stack Overflow to classify posts according to architectural lifecycle stage (analysis, synthesis, evaluation, ...). We focus primarily on design topics and their relationship to source code artifacts.

Other artifacts beside question and answer posts are relevant. Viviani et al. [39] attempted to recover design information from the social artifacts surrounding software engineering. Viviani et al. focused on developer discussions that take place on GitHub pull requests, as concerns related to design are often raised in those discussions.

Most closely related to our work is that of Mohamed Soliman and his co-authors. They have looked at using Stack Overflow as an architecture knowledge (AK) repository: first, in [34], assessing the "suitability of capturing and reusing of AK for technology decisions

from SO posts", specifically middleware knowledge. Their expert respondent pool agreed their sample of Stack Overflow posts was architecturally relevant. Next, they build an ontology for architecture knowledge on Stack Overflow [33]. The ontology is similar to our codes in the tables above, but include more architecturally specific terms such as *component, connector, architecture pattern* and then apply that ontology to a query-based retrieval system [35]. This last paper showed that questions with specific design details were harder to retrieve.

The primary differences with our paper is that we look more broadly at all design topics and use a semi-supervised approach to guide the topic model. We also use the term 'design', as opposed to their focus on architectural terms. We suspect design-related discussions may be more frequent, but these two terms do have considerable overlap. On Stack Overflow there are currently 15,650 questions tagged '[architecture]', out of the 1.5 million or so we consider in our dataset (which includes the '[architecture]' tag).

3 RQ1: EXTRACTING DESIGN TOPICS

We filter Stack Overflow questions using design-related tags; in RQ1.1 we inductively code that subset and then in RQ1.2 apply unsupervised LDA [9] and semi-supervised Anchored CorEx [15, 31, 38]. In Section 4, we then use CorEx to characterize the overall discussions on Stack Overflow.

3.1 Creating the Dataset

We build our dataset from Stack Overflow, which is a question and answer forum pertaining to programming with question commenting, question tagging, upvotes, and more. See [3] for more background on mining Stack Overflow. Filtering the design discussions from Stack Overflow data dump such as SOTorrent [3] is a challenging task, and is heavily dependent on the original user label for tagging of posts. Mahadi et al. [25] used 10 software-design related Stack Overflow tags (viz. "design-patterns", "software-design", "class-design", "design-principles", "system-design", "code-design", "api-design", "language-design", "dependency-injection" and "architecture") to identify design related discussions on Stack Overflow [25].

Tag Filtering—Stack Overflow follows a community-driven question and answer framework, wherein details such as appropriate creation and use of tags are monitored by expert users. The number of tags that can be assigned to a topic are limited to 5. These rules ensure that the tags that are assigned to a post remain germane to the topic of discussion.

To extend the dataset by Mahadi et al. [25], we use related tag analysis to identify tags and sub tags that are related to the already identified design tags. Furthermore, to include design related tags that were missed by related tag analysis, we query the tag description (identified by PostTypes: TagWikiExcerpt and TagWiki) for the keyword "design" and having more than 100 posts. We then manually analyze the resultant list of tags to remove tags that are not related to software design related posts. We avoid domain-specific tags such as `wordpress`, since such tags contained both design-related and non-design-related posts. We also exclude tags corresponding to graphic design and UI design (e.g., `css`).

Round	SO
Round 1	0.22 (fair)
Round 2	0.17 (slight)
Round 3	0.34 (fair)

Table 1: Inter-rater agreement [22] on Design Topic identification from codes in Table 2.

This systematic identification of design tags produced a tagset of 61 tags for *Stack Overflow* corresponding to software design. These tags are then used for identifying design-related questions and their corresponding answers which form a part of our extended dataset. The list of design-related tags and our dataset is available in our replication package, https://doi.org/10.5281/zenodo.5885783.

The extended design discussions dataset contains over 1.5 million design-related questions and answers. Pre-processing is performed on the posts to remove styling tags, code blocks and whitespace, to make the dataset easier for a human labeller to perform manual analysis and labelling.

Only question and answer type posts are retained in the dataset, along with their titles and associated tags. The dataset for Stack Overflow was obtained using the *SOTorrent dataset* [3], as of December 2020. We include the dataset and details on data format in our replication package.

3.2 RQ 1.1 Inductive Coding

3.2.1 Methodology. We use an inductive coding approach [27] to categorize the posts. We did not have a predefined coding guide, so we begin by having two authors, graduate students each with several years of commercial software development experience, independently label the posts, looking at title, question, and answers, alongside the post tags. We randomly sample 50 questions each for SO from our sampling frame (tag filtered dataset) for each round (3 total) of inductive coding. In each round, after reviewing 50 randomly sampled discussions independently, the coders met in a coding agreement session to discuss codes and look for themes. A third coder resolved disagreements. We maintain a coding dictionary for the set of codes identified. Table 2 shows our coding dictionary and relative code frequency out of the 150 questions coded for Stack Overflow. After three rounds of coding, reconciliation and agreement (50 posts per round; 150 posts in total), since no new labels were identified, the set of labels were finalized and the 150 posts were re-labelled using these finalized labels. We provide the labeled and coded dataset in our replication package. Table 1 reports the agreement kappa scores, for the Design Topic labels assigned. We discuss more on agreement in the results section.

Even after multiple rounds of coding and alignment, the agreement scores only showed small improvement. Therefore to identify the reasons for low agreement while manually classifying design related posts we perform an additional round of coding 75 Stack Overflow posts, with 3 authors labelling 50 posts each with a 50% overlap with each of the other coders. In this round of coding the coders were instructed to provide their top 2 labels for each of the posts, for analyzing topic overlap. We use the same coding dictionary that was arrived at after the previous 3 rounds of coding. We

Table 2: Inductive, manually assigned design-related codes. Counts are # questions (out of 150) associated with that code.

Codes	Description	SO
Other Design Questions	Discussion around efficiency, security and various best/recommended practice design choices.	31
OOP Design	Discussion around designing of classes and objects along with their associated attributes and methods.	19
Dependency Management	The various challenges of management and insertion of dependencies.	18
Architecture	Better ground up design of multiple systems/modules and their interaction/integration.	10
Design Patterns	General reusable solution to some commonly occurring software design problem e.g. GoF patterns.	10
Language Design	Challenges of language design and the discussion around new language constructs.	8
Data Storage Access Design	Design choices around selection of appropriate data structure/iteration strategies for storage/representation/accessing data.	7
Test Design	Design of test cases and their challenges.	4
Design Principles	Software design principles that mainly focus on the maintainability aspects of quality.	4
API Design	Design choices and challenges involved in developing APIs .	2
Not Design	*The question did not pertain to design despite its tag.*	37

also captured disagreement source, after examining the reasons for disagreements.

3.2.2 Results for Inductive Coding. Design is notoriously context-specific and somewhat ephemeral. We therefore expected to have challenges getting agreement on labels for these discussions. Table 2 lists our codes, short definitions, and frequency of occurrence. The more complete set of definitions is available in the replication package. We struggled to get good inter-rater agreement, as measured by kappa score, on the codes we assigned (Table 1).

In general the inductive coders were able to differentiate **design** and **non-design** posts well with a substantial agreement (Cohen's Kappa: 0.68)[22] prior to the alignment step. These results suggest that given a Stack Overflow post, it was relatively easy to identify whether it corresponded to Software Design.

Coder alignment for Design Topics (i.e., more refined subcategories of Design) proved more challenging. While the coders agreed on whether a discussion artifact corresponded to software design, they were only able to reach a fair agreement, for Stack Overflow, despite their background as professional software developers with experience in software design.

Even after three rounds of inductive coding and alignment, the software design topics identified overlapped and did not provide good separation in terms of inter-rater agreement. We did an additional round of coding to identify why this might be. We expand on this along with the percentage of all disagreements where this reason was cited (% Disagreements). Disagreement occurs when neither of the two assigned labels match for the two coders labelling it and each post may have multiple reasons for disagreement.

3.2.3 Challenges with Manually Labeling Design Topics. **Overlapping topics (37.5 % of disagreements).** The coding dictionary had multiple possible answers for a given question. For instance, *decorator pattern* is a GoF pattern but could be used as a solution for ensuring the *open-closed principle*, this represents a Design pattern-design principle overlap.

Questions asked were too basic (29.2%). Some questions caused confusion when the question was too simple. For instance, a Stack Overflow question[2] discussing the basics of designing a utility class was labelled by the 2 coders as being *OOP Design* and *Not Design* respectively.

Referred to different sections of the post (29.2%). The coders cited different parts of the post for justifying the different labels that were assigned. It is complex to summarize a complex question with one or two topics.

Judged error handling/bugs differently (20.8%). Some questions seemed more like specific bug issues than design questions. For instance, a Stack Overflow question[3] discussing handling errors with respect to project dependencies was labelled by the 2 coders as being *Dependency Management* and *Not Design* respectively.

Question too focused on implementation (8.3%). Some questions seemed to focus on implementation and not design. For instance, a Stack Overflow question[4] asking the correct implementation of a database query was considered by one of the coders as being too specific to be considered as software design, where as the other coder labelled it as *Data Storage Access Design* since it discussed efficiency and performance of the implementation. Note that here the user *did* consider it design.

Confusion between architecture and design (4.2%). The level of abstraction of the software design question might be either Design or Architecture. For instance, a post discussing some aspect of object oriented design at the project level may be labelled by the coders as both *OOP Design* and *Architecture*.

3.3 RQ 1.2 Topic Modelling

If the inductive approach is hard to scale (human labeling is expensive) and prone to disagreements, an unsupervised approach that relies solely on statistical models of the text might be more suitable. An unsupervised (and *semi supervised*) approach to recovering latent topics in text is topic modeling. It has the benefit of being highly scalable (e.g., can handle all of Stack Overflow) and has a long history in SE [1]. We compare the conventional unsupervised

[2] https://stackoverflow.com/questions/30019005/
[3] https://stackoverflow.com/questions/23663997/
[4] https://stackoverflow.com/questions/7711432/

approach, LDA, to a newer semi supervised approach, CorEx, which uses anchoring. Weak supervision (or semi-supervised) refers to the fact that users do some labeling, but much less than inductive coding.

3.3.1 Latent Dirichlet Allocation (LDA). We use the LDA algorithm [9] as implemented in Mallet[5] to automatically identify topics of design-related posts on Stack Overflow.

We grid search for the number of topics between the limits we define as max (70) and min (2) topic numbers to retrieve the number of topics with highest coherence score. These seem like a reasonable prior to apply based on studies in SE using LDA (such as [4]).

The data set of Stack Overflow consisted of 227,282 design questions with a vocabulary size of 259,167. However, removing stop words, short tokens before and after bi-gram and tri-gram lemmatization resulted in a vocabulary size of 213,329. We wanted to further narrow down the vocabulary size to improve the model. So, we added high-frequency words to the stop words list and filtered out words that occurred in fewer than 10 documents, or more than 50% of the documents. This reduced the vocabulary to 30,457 words.

We used Mallet's optimization technique to determine the hyperparameter for the optimal number of topics for the given design discussions data set by comparing their c_v coherence scores. Seven topics are optimal for the Stack Overflow dataset, with a coherence score of 0.56.

3.3.2 BERT and S-BERT. To see if we could improve on vanilla LDA, we used Bidirectional Encoder Representations from Transformers (BERT) [11] and Sentence Transformers [30] to perform semantic textual similarity (STS). We used Adam Optimizer [21] and mean squared error loss function for the encoding process. This resulted in coherence scores of 0.54 and 0.59 for BERT and S-BERT respectively. We noticed that the deep learning models, being significantly more complex than LDA model, creates more noisy features to detect similarities and hence, the coherence scores are not better than the vanilla LDA model, so we did not proceed further. The sentence embeddings from the pre-trained BERT and S-BERT without fine-tuning have been found to poorly capture semantic meaning of sentences [23].

3.3.3 c-TF-IDF and Anchored CorEx. We compared LDA, which is unsupervised, to a semi-supervised method for topic modelling, by using a combination of Class Based Term Frequency-Inverse Document Frequency(c-TF-IDF) and Anchored Correlation Explanation (Anchored CorEx) [15, 31, 38]. The idea is to use domain knowledge (supervision) to anchor the topics semantically. We use our inductive coding results (Table 2) as topic labels. We also must determine what useful anchor terms would be for each of those labels. We do this with c-TF-IDF.

Class Based Term Frequency-Inverse Document Frequency(c-TF-IDF). We use c-TF-IDF to identify the top 30 informative terms (i.e., candidate anchor terms for CorEx, below) for each of the classes in the inductively coded dataset. The top 30 terms identified per class provides the latent representation of the classes as terms that can strongly distinguish a class. c-TF-IDF is a TF-IDF like method for feature representation, where each document represents a class. The c-tf-idf for every term t and a class document d is given as

$$c - tf - idf(t,d) = tf(t,d) * (log\frac{1+N}{1+df(t)} + 1) \quad (1)$$

where $tf(t,d) = f_{t,d}/\sum_{f' \in d} f_{t',d}$ is the term frequency of the term t in the class document d with $f_{t,d}$ term count, N is the number of class documents, and $df(t)$ is the document frequency of the term t (the number of class documents containing the term t).

We preprocess the inductively coded Stack Overflow dataset by performing lemmatization, uni-gram and bi-gram tokenization, and generate the class documents by merging the posts based on their assigned codes. The standard scikit-learn[6] TfidfVectorizer on the class documents to generate the c-tf-idf values which are then ranked to extract the top 30 informative terms per class.

The top 30 informative terms are then manually analyzed and filtered to retain only meaningful terms (for that label) and to eliminate spurious correlations. Table 3 presents a subset of the inductive labels, showing the top 30 terms identified per class and the meaningful terms retained to be used as anchor words for Anchored CorEx (in blue). For example, in Topic 8 (Not-Design) words such as 'entity' or 'module' are clearly design related, and thus we do not choose them as meaningful representations for the Not-Design topic. On the other hand, words like 'songwriter' reflect the sparsity of the dataset and, while not design, nonetheless seem unlikely to anchor topics properly (in our design exploration, anyway).

Anchored Correlation Explanation (Anchored CorEx). Anchored CorEx [15, 31, 38] is a semi-supervised approach for topic modelling, that allows incorporating domain knowledge in topic modelling by specifying anchor words. We use the terms identified through c-tf-idf as weak supervision anchor words that seed and impose semantics on latent factors while performing Anchored CorEx. Anchored CorEx involves optimizing the equation:

$$Maximize \quad TC(X;Y) + \beta \sum_{i,j \in R} I(X_i, Y_j) \quad (2)$$

where $TC(X;Y)$ represents total correlation with the objective of constructing latent variables Y that best explain the multivariate dependencies in data X (set of Stack Overflow questions), and $I(x,y)$ represents maximizing mutual information between the anchor terms X_i and the latent factors Y_j.

We pre-process the dataset, performing stop word removal, lemmatization, and unigram, bigram tokenization. We then use the Anchored CorEx implementation[7] and perform simple grid-search hyperparameter optimization for coherence scores (C_v) by adjusting for anchor strength, a measure of how strong the anchor term is. In case of overlapping (a spurious correlated term) we can add the overlapped term to the appropriate term as a new anchor. We find that an anchor strength of 6 provides the highest coherence score of 0.55 and a total correlation score of 60.11.

3.3.4 Results for Topic Modelling. Topic modeling, as a classification approach, is faster and vastly more scalable than inductive (human) labeling. We use the the coherence score C_v as a metric

[5] http://mallet.cs.umass.edu

[6] https://scikit-learn.org

[7] https://github.com/gregversteeg/corex_topic

Topic Id	Topic	Terms
2	architecture	datablock, content, dll, class, assembly, abstraction, use, version, web, application, delegate, app, interface, block, model, datum, service, object, need, assembly version, concern, not, menu, file, wpf, app delegate, main app, method, system, architecture
5	design-pattern	node, command, view, class, pattern, object, render, update, method, implement, scene, not, strategy, scenegraph, scene graph, state, throw, rendergraph, opengl, issuer, interceptor, command pattern, pipeline, model, type, graph, singleton, event, submit, transition
6	design-principles	deck, card, design, structured, structured design, focus, design principle, card deck, workspace, card collection, middle, model, principle, user, object, parent instance, middle entity, deck model, collection, author, software design, instance, reference, bind, scope, hierarchy, parent, module, not, new, principle
8	not-design	not, class, method, use, error, try, server, file, module, verilog, model, event, php, entity, songwriter, image, code, work, window, table, js, need, variable, constructor, self, database, framework, text, want, self songwriter, syntax, validation, exception
9	oop	class, method, not, object, constructor, employee, vector, argument, need, static, superclass, function, base class, atm, subclass, instance, static method, inherit, base, use, mutable, design, inheritance, create, value, input, php, object orient, call, code

Table 3: Top 30 representative terms for a selection of inductively defined design topics (Table 2), retrieved using c-TF-IDF. Blue are the selected anchor terms. Black terms are discarded as anchors. Red are anchor terms added to overcome topic overlap.

Table 4: Stack Overflow LDA topic names (manually assigned) with top 5 terms

Topic Name	Top 5 Terms
oop	object, property, variable, instance, field
concurrency pattern	thread, call, task, request, resource
class/interface implementation issues	class, method, function, type, interface
error resolutions	module, file, dependency, work, component
other design	return, work, element, item, iterator
design patterns and principles	pattern, thing, person, language, feature
databases and architecture	datum, service, user, model, event

Table 5: Stack Overflow CorEx topic names and top 5 associated unigrams/bigrams, exclusive of anchors.

Topic Name	Associated Terms
api-design	good, thing, think, case, need
architecture	web, web application, framework, project, net
data-storage-access-design	view, data, controller, entity, sql
dependency-mgmt	service, container, ioc, inject dependency, resolve
design-pattern	singleton class, singleton pattern, factory pattern, pattern use, factory
design-principles	design pattern, software design, responsibility, good design, drive
language-design	create object, object not, object class, object create, orient
not-design	throw, get error, error message, throw exception, try
oop	static method, object orient, constructor, interface, class method
other-design-questions	ddd, wait, thread safe, domain object, locking
test-design	unit, mock, testing, test code, test class

for evaluating the topic models since it is strongly correlated with human labelling and is reliable for topic coherence evaluation [32].

Table 4 shows the results of LDA topic modelling with our manually assigned topic names. While LDA provides an acceptable coherence score (0.56) for topic modelling [32], it does so by optimizing for coherence based on the word vector representations of the data. However, the LDA analysis results in only 7 design topics, which seems low given the breadth and scope of software design topics.

Table 5 shows the results of Anchored CorEx topic modelling and the top 5 terms identified per topic. We do not report the anchor terms (i.e., the actual word list would be Table 5 + the terms in blue in Table 3). We note that for some of our topics, such as [api-design] and [language design] topics, our labeling dataset had relatively few training examples (e.g., only 2 questions for [api-design] – cf. Table 2). This explains some of the strange terms (like 'good').

Anchoring results in CorEx topics that are interpretable, because they can be tied back to the domain knowledge captured in the inductive coding, while having coherence scores for Stack Overflow dataset that are on par with LDA topic modelling (CorEx: 0.55, LDA: 0.56).

Table 6: Topic Modelling precision (P) and recall (R) metrics. Parentheses indicate the LDA/CorEx topic name we matched on.

SO Tag	LDA (lda-topic-name)	Anchored CorEx (corex-topic-name)
design-patterns	**P: 0.25** R: 0.17 (design patterns and principles)	P: 0.10 **R: 0.74** (design-pattern)
solid-principles	**P: 0.013** R: 0.27 (design patterns and principles)	P: 0.004 **R: 0.84** (design-principles)
oop	**P: 0.39** R: 0.22 (oop)	P: 0.19 **R: 0.71** (oop)
architecture	**P: 0.16** R: 0.35 (databases and architecture)	P: 0.05 **R: 0.77** (architecture)

Comparing topic model approaches. Which of the two topic modeling approaches is better, i.e., which one better captures latent design discussions? One way to answer this question is to see how well each named topic captures *existing* annotations. We use precision and recall metrics to evaluate the performance of the two topic modelling approaches against the user-created tags on each question from Stack Overflow. A 'true positive' is when a named topic from either LDA (column 1 of Table 4) or CorEx (column 1 of Table 5) matches a user-assigned tag on Stack Overflow. In some cases we mapped a tag from Stack Overflow users (such as "solid-principles") to our named topics in LDA and CorEx (such as "design-principles" (CorEx) or "design patterns and principles" (LDA)).

Table 6 explores the precision and recall results for a selected set of topic names. The parentheses list the mapping between the SO tag and the topic modeling name. We see that CorEx generally has much higher recall (finds more true positives) but at the expense of lower precision (more false positives). We explore reasons for this in Section 5.

4 RQ2: LEVERAGING DESIGN TOPIC KNOWLEDGE

With a suitably performant topic extraction approach, in our case, CorEx, there are several questions one can attempt to answer. We describe two of them in this section. One concerns the types of topics discussed on Stack Overflow, and the other looks at how GitHub and Stack Overflow design topics relate.

4.1 What Design-Related Topics are Discussed on Stack Overflow?

Given the challenges with classifying design related posts using inductive coding and the low agreement scores, and the advantages with using a semi-supervised model of classification we use CorEx Topic Modelling for classifying design related posts on Stack Overflow.

Table 7: CorEx topics. Counts reflect number of SO questions having some discussion around that label.

CorEx Label	SO frequency
oop	89,270
language-design	88,796
data-storage-access-design	74,371
architecture	73,494
not-design	65,579
api-design	63,972
design-pattern	54,403
design-principles	47,706
dependency-mgmt	38,045
other-design-questions	37,131
test-design	25,569

Stack Overflow questions were spread widely across the topics from Table 2, and focus more on specific questions related to OOP (19/150) and dependency management (18/150).

Despite being filtered for design-related tags, we found 24.6% of Stack Overflow posts were not design-related. By contrast, examining non-tagged artifacts such as code comments or issue discussions (as in [25] or [39]) found the inverse: 75-90% of the Stack Overflow posts were not design-related. This shows the effectiveness of using the tag filtering approach. For instance, Stack Overflow's tag description for the tag "iterator" says the tag is for questions related to the Iterator design pattern (a Gang of Four [16] Behavioural design pattern). But this tag was often found to be incorrectly used[8] for tagging loop implementation issues, and we labelled those questions as "Not Design".

Table 7 shows the number of questions associated with each of the CorEx topics. Note that 'not-design' again figures fairly highly, despite our tag filtering approach; either these questions are mistagged (e.g., with 'iterator'), or some of the latent topics are incorrect. Figure 1 presents the co-occurrence for the various CorEx terms in the Stack Overflow dataset labelling. Here we observe that terms such as 'dependency-mgmt', 'not-design' and 'test-design' have a comparatively lower co-occurrence with other design labels, suggesting that the Stack Overflow posts tagged with these labels tend to be more focused and have very few latent terms associated with the other design topics, the hoped-for result for 'not-design'. We also observe a strong spurious co-occurrence between 'oop' and 'language-design', which may be explained due to the unavailability of sufficient training samples.

4.2 Linking GitHub Data to Stack Overflow Design Topics

To explore what design related Stack Overflow discussions are referenced in open source projects, we examine the *PostReferenceGH* and *GHCommits* tables available as a part of the *SOTorrent*[3] dataset. The *PostReferenceGH* and *GHCommits* tables provide references of the Stack Overflow questions, answers, and comments that are made in source code (i.e., code comments) and commit messages,

[8] e.g., https://stackoverflow.com/questions/38024554/

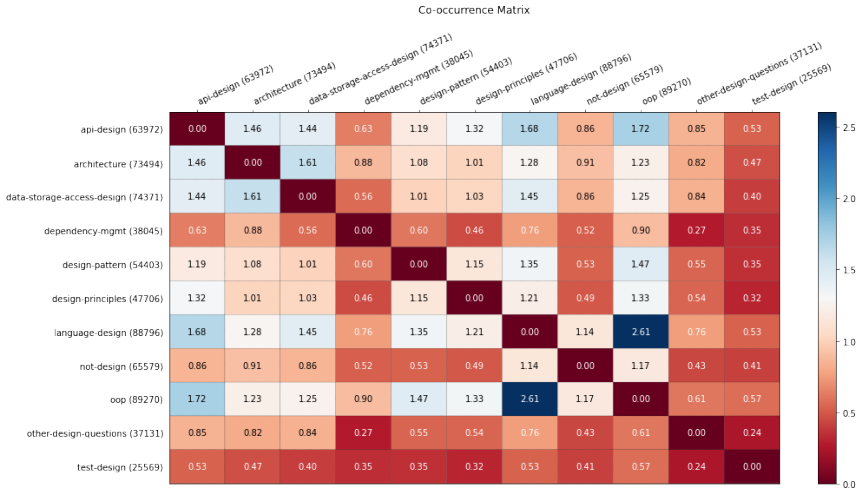


Figure 1: Co-occurrence matrix for the CorEx topics in Stack Overflow dataset. Values indicate the percentage co-occurrence of labels of all Stack Overflow design related posts. Higher (more blue) = more likely to co-label a SO question.

respectively. We identify a subset of the Stack Overflow extended design discussion dataset posts that were referenced in GitHub repositories using these mapping tables and analyze them using both the tags used by the original poster (or modified by site moderators) as well as the labels generated by our Anchored CorEx model. We ensure that we avoid duplicate rows due to forking for commits, but not for code comments, as the forks may change the comment.

We used our CorEx topics to characterize the types of questions GitHub projects referred to. We found 23,634 source code references and 2,251 (221 after grouping by *CommitId* and *PostId*) commit message references in GitHub to Stack Overflow design related posts (the posts tagged with a word from our tag set).

Table 8 presents the distribution of the Stack Overflow posts referenced in GitHub as classified by the CorEx model.

Stack Overflow design discussions form 0.36% and 1.13% of all source code and commit message references respectively, a high percentage of references given these posts were identified using only 61 design related tags as compared to the 20,000+ tags on Stack Overflow that have at least 100 posts.

According to the user defined tags, discussions relating to model-view-controller and oop tagged posts were most referenced in GitHub source code and commit messages respectively. Since CorEx looks for the presence of multiple labels per reference, we observed that most of the source code references shared the architecture label, while most of the commit messages shared the data-storage-access-design label. Further analysis of some the references under these dominant categories reveals some of the ways in which GitHub projects utilize them.

Table 8: Distribution of Stack Overflow references in GitHub as classified by the CorEx model. The count refers to the number of code comments or commits messages that link to a design question with the associated CorEx topic in SO.

CorEx Label	Source Count	Commit Count
api-design	21360	1974
language-design	19830	1006
architecture	13928	1070
oop	10990	993
not-design	9793	1220
design-pattern	9267	804
data-storage-access-design	6196	1340
test-design	5553	678
design-principles	4232	611
other-design-questions	4111	1337
dependency-mgmt	2372	562

For example, a GitHub commit message[9] referenced Stack Overflow discussion on providing synchronized access to data model using double checked locking: "*...need to make sure the initialization of the content provider component is synchronized... accepted answer here: http://stackoverflow.com/questions/5717090/double-checked-locking-in-android/5717977*" to corroborate and to document the strategy used in the commit. CorEx labeled this SO post as [data-storage-access-design, other-design-questions] where 'other-design-questions' corresponds to locking strategy used. The Stack Overflow poster

[9]https://bit.ly/3FOUZCv

tagged it [java, android, virtual-machine, dalvik, double-checked-locking].

A source code reference in a logging library[10] commented " ... *This class was retrieved from: http://stackoverflow.com/questions/33364070/ ...*" where the Stack Overflow post presents a design strategy for implementing singleton design pattern and inheritance from abstract classes. CorEx classified this as [design-pattern,oop] and the poster tagged it as [python, singleton, abstract-class, metaclass, abc].

Since 1% of GitHub-linked discussions are design-related, we could use this dataset for both generating documentation (e.g., on-demand as someone reads the code) and adding documentation sources for writers. One use case, for instance, could be to retrieve Stack Overflow posts given source code similarity. Something like this is shown by Xu et al. [40] for code generation.

Soliman et al. [34] found a similar phenomenon with architecture knowledge for middleware. It may be possible to add automation to this and suggest relevant Stack Overflow style discussions as source code comments. Such simple links would be considerably simpler to document than the entire design choices, much like naming a class SingletonX implicitly links to the Singleton design pattern (for those who have learned that pattern).

5 DISCUSSION

5.1 Un-/Semi-/Fully-Supervised Approaches

We used both inductive coding (a manual approach), and LDA/-CorEx (unsupervised and semi supervised approaches). A fourth topic set is the user-assigned *tags* on the Stack Overflow question. All approaches label the content of the question with respect to design. Table 9 shows the Stack Overflow question title, the manual, inductive code we assigned, the user-defined tags (where *user* is the asker of the question on Stack Overflow and possible editors of those tags), and finally, the set of CorEx topics associated. Table 10 captures our summary of the benefits of each approach.

We leveraged the existing tags on SO to generate our design Q&A dataset. We then inductively coded this dataset. One might ask why the coding process was even necessary, given the users had already defined tags. Our inductive coding process is high-effort, but we showed that many posts tagged as [design] were in fact not design.

Furthermore, tagging on SO is contentious, including for the reasons Barua et al. [5] mention: minor syntactic differences, tag evolution, and tag semantics (e.g., design pattern vs design principle may be perceived differently by different people). More importantly, manual coding allows us to thematically group the posts and identify broader patterns (such as the motivation for asking the question). Furthermore, we were able to use this inductive set as input for the CorEx approach which is much more scalable.

On the differences between LDA and CorEx. We expand on the differences captured in Table 10. Is it a problem that CorEx has higher recall but lower precision, in general?

P/R is a tradeoff: it may be better in some applications to get all the relevant questions instead of focusing on precision, and a balanced approach may not be appropriate [7]. Our evaluation is exploratory; for example, it is possible users are missing a label which should really be there (i.e., users may be mistaken)

Our definition of a true positive may be overly constrained. Just because our two coders tagged a question as 'architecture' does not imply the original question was about architecture; conversely, the user tag might indeed be architecture, but at a lower abstraction level (such as 'hexagonal'). Finally, our inductive coding, and design labeling in general, is imperfect at the human level, which suggests tools will also struggle without better definitions. We see CorEx as providing a useful way, with anchoring, to more precisely define design concepts, training the topic model with less common or generic topics, which LDA suffered from. Another difference is in how topics are named. This is a tricky and subjective aspect of LDA [20]. With CorEx, these names come from the anchor set a priori and there is no labeling topics post-hoc.

One of the advantages of LDA-style approaches is that they are objective (to some degree; choice of hyper-parameters and dataset order affect the reliability [1]). Inductive coding, although rigorous [27], depends on the skill and experience of the coders. Yet we have shown that *some* supervision greatly increase topic relevance. This suggests that a balance is important. In this paper, some aspects were objective: (randomly) selecting sample data; running LDA; selecting hyper-parameters; and running CorEx. Some were subjective: inductively coded data; naming LDA topics; selecting anchor words using c-TF-IDF. We argue that **this balance is inherent in SE domains, such as design, which are not objectively defined themselves**.

For example, while selecting anchor terms, we cannot simply remove terms, because those terms might well be relevant, and it is our knowledge of the topic that is lacking. On the other hand, as Agrawal et al. have reported [1], repeatability and stability in topic modeling are important as well. Ideally the topics CorEx identified in this study would be relevant in another study (on the same data).

Guidelines for semi-supervised topic modeling in SE. Our experiences lead us to suggest the following guidelines.

(1) Clearly specify criteria to avoid topic overlap. To ensure clear topic separation, merge highly overlapping topics. Use of hierarchical topic modelling could also help topic separation.
(2) Filter out posts/documents using (a) experience criteria for the author of the post; (b) exclusion criteria for code snippets and error logs; (c) exclusion criteria for questions that do not have a selected answer
(3) Clearly define a goal for the resultant topic model. This would help streamline the topic modelling activity and the definition of the anchor terms. For example, Soliman et al.'s investigation into middleware [34].
(4) Experiment with various sampling frames and document granularities (document, paragraph, sentence). A more granular post type, may improve precision, but may also require more efforts to scale to a large sampling frame,
(5) For design mining, consider the scale of the design solution being looked at (method level, file level, program/architecture level).

[10] https://bit.ly/3fH1AnK

Table 9: Example SO questions, and approaches to identify latent topics

Question Id & Title	Inductive Code	User Defined Tags	CorEx Topics
11242144: *Optimal TPL Dataflow Design*	Architecture	c# architecture asynchronous concurrency tpl-dataflow	api-design,architecture,design-principles,language-design
34258867: *How factory pattern works*	Design Patterns	php oop factory	design-pattern,language-design,oop

Table 10: Four approaches to categorizing latent design topics.

Approach	Pros	Cons
LDA	Efficient, tunable. Assign many (all) topics to a question.	Potentially unstable; topics hard to explain; sensitive to dictionary.
CorEx	Improved recall on LDA, efficient, scalable. Allows for expert input (semi-supervision) making topics more interpretable.	Lower precision. Possibly unstable: subject to bias from supervision.
Inductive Coding	Qualitative and expertise-based, handles nuance.	Not scalable; sensitive to sample; coder agreement low.
Tags	Crowd-sourced, specific.	Edited by others; not all design.

(6) While performing anchoring for semi-supervised topic modelling, (a) experiment with various seed words, anchor strength and topic numbers; (b) after every iteration manually access overlapping terms and anchor them as needed; (c) use strategies such as c-TF-IDF to glean potential anchor words; (d) refer to [15] for anchoring strategies for topic separability, representation, and aspects.

5.2 Potential Limitations

Internal Validity: We filter our dataset using a set of design tags; therefore, if a design discussion uses a tag not in that set, we would miss this as a potential source of data. To mitigate this we conducted several rounds of tag set expansion (see Section 3.1). Barua et al. [5] highlight some issues, including problems with minor, syntactic changes in tags that do not affect semantics, and tags which are quite generic. In our case, the tag "design-pattern" seems to occur even when the question is not specific to a known pattern like Singleton. Tags also evolve, and indeed, Stack Overflow moderators recently completed a major refactoring of the previous '[Design]' tag, re-tagging posts with more specific tags.

Our anchored/supervised topic model approach using CorEx relied, for some topics, on fewer than 10 "documents" (SO question/answer sets). This means the topic model had limited supervision on these topics and this might limit the applicability of the matched questions. However, we reported recall and precision to show how well the anchoring worked, even with limited training.

Since we had poor agreement between coders, one might argue our coding dictionary was poorly constructed. We believe this was not the main problem, and that rather, the problem is that design in software is poorly defined. Devising a coding scheme that can delineate design concepts is tricky. And yet many of the SO questions we examined were clearly in one category or the other, suggesting there is a difference.

Our recall and precision true positives were based on alignment between user tags and our topic names. This was a reasonable choice to explore the differences but a more sophisticated metric will be more informative.

External Validity/Sampling Frame: The objective of our paper is to identify various software design related questions asked and the software design related challenges faced by developers, as discussed on public question and answer forums. The theoretical population that this paper is concerned with are the design topics discussed (i.e., in a question or answer) by software developers on public Q&A forums. For this we purposively choose the design related questions asked (using our tagsets) on Stack Overflow as our sampling frame. Our results should generalize to our chosen sampling frame. Other studies utilizing a different sampling frame may find design topics relevant to the type of the forum being studied. We examined GitHub Discussions as another potential source, but found there were very few discussions about design.

Manual coding: The design mining data relies on a limited set of human labelled data (or makes the possibly invalid assumption that the tagging in Stack Overflow reflects real design). However, we reached code saturation, and we show that the LDA topics and the human codes are quite similar. Thus, we do not believe we missed any substantive codes.

Ethical considerations. Stack Overflow is a public website and content created there is acceptable for research according to the Stack Overflow licence[11]. GitHub defines Acceptable Use Policies we adhere to.[12] However, while our study looked at publicly available archival data protected by law, and reports on non-personal information, ethical considerations including privacy concerns and possible harm, such as reputational damage, are relevant, should we expand the scope of the research to, for example, consider individual questions. At that point, informed consent should be obtained [17].

6 CONCLUSION

Software design is an inherently subjective, yet important, aspect of software development. After building a design-related Stack Overflow dataset, we used an inductive coding method to identify the design-related topics/areas discussed on Stack Overflow. We found

[11]https://stackoverflow.com/legal/terms-of-service#access
[12]https://docs.github.com/en/github/site-policy/github-acceptable-use-policies#5-information-usage-restrictions

it difficult to get coder agreement on certain design topics and list some possible reasons. We then used two scalable techniques, LDA and CorEx, to characterize the topics in the dataset. We found the semi-supervised CorEx approach was more interpretable and produced more coherent topics. That allowed us to characterize what design topics occur in Stack Overflow, and how comments on GitHub refer to those topics. We then discussed the differences we found in characterizing latent design topics with four approaches, and listed some guidelines for other studies using semi-supervised topic modeling. Our experience using the semi-supervised CorEx approach leads us to believe that approaches like CorEx that combine domain knowledge and scalability are key for analyzing SE text repositories, particularly those with subjective latent topics like design.

ACKNOWLEDGMENTS

Omitted for double blind.

A REPLICATION PACKAGE

The Stack Overflow design discussions data set used for this study, the LDA model generated and the manual coding sample set used for each topic generated is provided in our replication package at https://doi.org/10.5281/zenodo.5885783.

REFERENCES

[1] Amritanshu Agrawal, Wei Fu, and Tim Menzies. 2018. What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology* 98 (jun 2018), 74–88.

[2] Mubashir Ali, Husnain Mushtaq, Muhammad B Rasheed, Anees Baqir, and Thamer Alquthami. 2021. Mining software architecture knowledge: Classifying stack overflow posts using machine learning. *Concurrency and Computation: Practice and Experience* (March 2021). https://doi.org/10.1002/cpe.6277

[3] Sebastian Baltes, Lorik Dumani, Christoph Treude, and Stephan Diehl. 2018. SOTorrent: reconstructing and analyzing the evolution of stack overflow posts. In *Proceedings of International Conference on Mining Software Repositories*. ACM, 319–330. https://doi.org/10.1145/3196398.3196430

[4] Abdul Ali Bangash, Hareem Sahar, Shaiful Chowdhury, Alexander William Wong, Abram Hindle, and Karim Ali. 2019. What do Developers Know About Machine Learning: A Study of ML Discussions on StackOverflow. In *International Conference on Mining Software Repositories (MSR)*. Montreal, QC, Canada, 260–264. https://doi.org/10.1109/MSR.2019.00052

[5] Anton Barua, Stephen W. Thomas, and Ahmed E. Hassan. 2012. What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering* 19, 3 (Nov. 2012), 619–654. https://doi.org/10.1007/s10664-012-9231-y

[6] Len Bass, Paul Clements, and Rick Kazman. 2012. *Software Architecture in Practice* (3rd ed.). Addison-Wesley Professional.

[7] Daniel M. Berry, Ricardo Gacitúa, Peter Sawyer, and Sri Fatimah Tjong. 2012. The Case for Dumb Requirements Engineering Tools. In *International Conference on Requirements Engineering: Foundation for Software Quality (REFSQ)*.

[8] Tingting Bi, Peng Liang, and Antony Tang. 2018. Architecture Patterns, Quality Attributes, and Design Contexts: How Developers Design with Them. In *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*. 49–58. https://doi.org/10.1109/APSEC.2018.00019

[9] Joshua Charles Campbell, Abram Hindle, and Eleni Stroulia. 2015. Latent Dirichlet Allocation. In *The Art and Science of Analyzing Software Data*. Elsevier, 139–159. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

[10] Rafael Capilla, Anton Jansen, Antony Tang, Paris Avgeriou, and Muhammad Ali Babar. 2016. 10 years of software architecture knowledge management: Practice and future. *Journal of Systems and Software* 116 (June 2016), 191–205. https://doi.org/10.1016/j.jss.2015.08.054

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Osama Ehsan, Safwat Hassan, Mariam El Mezouar, and Ying Zou. 2021. An Empirical Study of Developer Discussions in the Gitter Platform. *ACM Transactions on Software Engineering and Methodology* 30, 1 (jan 2021), 1–39. https://doi.org/10.1145/3412378

[13] Denae Ford, Justin Smith, Philip J Guo, and Chris Parnin. 2016. Paradise unplugged: Identifying barriers for female participation on stack overflow. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. 846–857.

[14] Peter Freeman and David Hart. 2004. A Science of Design for Software-Intensive Systems. *Commun. ACM* 47, 8 (Aug. 2004), 19–21. https://doi.org/10.1145/1012037.1012054

[15] Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics* 5 (2017), 529–542.

[16] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. 1994. *Design Patterns:Elements of Reusable Object-Oriented Software*. Addison-Wesley.

[17] Nicolas E. Gold and Jens Krinke. 2020. Ethical Mining. In *Proceedings of the 17th International Conference on Mining Software Repositories*. ACM. https://doi.org/10.1145/3379597.3387462

[18] Ian Gorton, John Klein, and Albert Nurgaliev. 2015. Architecture Knowledge for Evaluating Scalable Databases. In *2015 12th Working IEEE/IFIP Conference on Software Architecture*. 95–104. https://doi.org/10.1109/WICSA.2015.26

[19] Hideaki Hata, Nicole Novielli, Sebastian Baltes, Raula Gaikovina Kula, and Christoph Treude. 2021. GitHub Discussions: An exploratory study of early adoption. *Empirical Software Engineering* 27, 1 (Oct. 2021). https://doi.org/10.1007/s10664-021-10058-6

[20] Abram Hindle, Neil A. Ernst, Michael W. Godfrey, and John Mylopoulos. 2012. Automated topic naming. *Empirical Software Engineering* 18 (2012), 1125–1155.

[21] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[22] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (March 1977), 159. https://doi.org/10.2307/2529310

A.7 Ethics - Session 1 - Email to Participants

Hi <PARTICIPANT-NAME>!

Thank you for expressing interest in our study!

My name is Cassandra Cupryk and I'm a master's student who is a part of a research team at the University of Victoria in Canada. I'm really interested in finding out whether a tool that has been built can help Computer Science Master's and Ph.D. students better review research papers.

Thus, we are currently looking for Computer Science Master's and Ph.D. students who are interested in improving how they review research papers. If you fall under this category, we would really appreciate the opportunity to:

(A) let Us observe you for about 3 hours in an online group setting while you:
   a. use the tool to review 2 research papers
   b. download the 2 reviews, in form of text files, of the 2 research papers to your device
   c. upload the content of the 2 text files to a survey
   d. express your opinion of the tool in the survey, and
   e. discuss your opinions of the study with other people

(B) let Us analyze your reviews of the 2 research papers and your completed survey

(C) let Us record your video and your audio for the duration of the study

(D) answer questions about whether you found the tool useful for reviewing the 2 research papers

(E) answer questions about what improvements could be made to the tool

We will openly publish the results so everyone can benefit from them but will anonymize everything before doing so. We will handle your responses and data confidentially. This is a purely academic research project with no commercial interests. As a master's student, I am required to conduct research as part of the requirements for a degree in Computer Science. The study is being conducted under the supervision of Dr. Storey. You may contact my supervisor at mstorey@uvic.ca if necessary.

If you're willing to participate, the study will be conducted on Friday, June 24th from 1pm to 4pm (PT). We estimate that this study will take about 3 hours of your time.

In addition, we are offering a $20 Amazon gift card to anyone who participates in the study.

If you are interested in participating in this study, please read, sign, and return the attached consent form.

After you send the signed consent form, you will be sent a calendar invitation via email for the study.

If you consider this email to be spam, I'm very sorry! There will be no follow-up to bug you.

Thanks, and have a wonderful day!

Cassandra Cupryk, ccupryk@uvic.ca

A.8 Ethics - Session 1 - Consent Form to be Signed by Participants

University of Victoria

Participant Consent Form Contextual Inquiry

Evaluation of the Empirical Standards Tool

We are Cassandra Cupryk and Margaret-Anne Storey from the Computer Human Interaction and Software Engineering Lab (CHISEL) at the department of Computer Science at the University of Victoria, Canada as well as Paul Ralph from the department of Computer Science at the University of Dalhousie, Canada. The purpose of the study is to determine whether a tool that we've built can help Computer Science Ph.D. and Master's students better review research papers.

Thus, you are being invited to participate in the study entitled "Evaluation of the Empirical Standards Tool". We would appreciate the opportunity to gather data on how you'd use the Empirical Standards tool to review research papers as well as if you found the Empirical Standards tool to be beneficial.

We would be delighted if you would be willing to:

(A) let Us observe you for about 3 hours in an online group setting while you:
   a. use the tool to review 2 research papers
   b. download the 2 reviews, in form of text files, of the 2 research papers to your device
   c. upload the content of the 2 text files to a survey
   d. express your opinion of the tool in the survey, and
   e. discuss your opinions of the study with other people

(B) let Us analyze your reviews of the 2 research papers and your completed survey

(C) let Us record your video and your audio for the duration of the study

(D) answer questions about whether you found the tool useful for reviewing the 2 research papers

(E) answer questions about what improvements could be made to the tool

This is a purely academic research project with no commercial interests. As a master's student, I am required to conduct research as part of the requirements for a degree in Computer Science. It is being conducted under the supervision of Dr. Storey. You may contact my supervisor at mstorey@uvic.ca if necessary.

We will openly publish the results so everyone can benefit from them, but will anonymize everything before publishing the results. Your responses will be handled confidentially. Please note that you are not obligated to participate in the study. If at some point during the study you want to stop, you are free to do so without any negative consequences and any data collected up to that point will be discarded.

Participants Selection

You are being asked to participate in this study because you have been identified as a Computer Science Ph.D. or Master's student who is currently looking to improve how they review research papers. You have also indicated that you are willing to assist with our study by agreeing to being audio and video recorded.

What is involved

If you consent to voluntarily participate in this research:

- You will be asked to participate in an online study in a group setting with other individuals.
- You will be provided the website URL for the Empirical Standards tool.
- You will be asked to use the tool to review 2 research papers and download the 2 reviews to your device and upload the 2 reviews to the survey
- You will be asked to complete a survey to express your opinion of the Empirical Standards tool.
- You will be asked to be a part of a discussion to express your opinion of the study.

Risks

There are no known or anticipated risks to you by participating in this research. The study will be completed in an online setting where other Computer Science Ph.D. and Master's students and study investigator(s) will be present.

Benefits

The potential benefits associated with the study include the development and improvement of a tool that can help Computer Science Ph.D. and Master's students better review research papers.

Voluntary Participation

Your participation in this research must be completely voluntary. If you do decide to participate, you may withdraw at any time without any consequences or any explanation. If you do withdraw from the study your interview responses will be deleted and will not be used in any form in our current (and future) study. We will also give you a $20 Amazon Gift Card after the conclusion of this interview. This $20 Amazon Gift Card will be given whether you finish the study or not, so you are not forced to finish the study just to get the incentive.

Confidentiality & Anonymity

The workshop is done within a group setting, so the researcher cannot guarantee confidentiality. However, after the study, your data and your information will be access restricted, and password protected, anonymized and will be kept confidential. The researchers are the only ones who will have access to this data. All of the data will be stored on a University of Victoria Microsoft Office 365 OneDrive Account. OneDrive is built on the Microsoft Office 365 hyper-scale, enterprise-grade cloud, which delivers advanced security and compliance capabilities. The data is encrypted in transit and at rest. All OneDrive files in the University of Victoria Microsoft 365 account are stored in Canada, however, some automated processing may occur outside of Canada. After conducting the study, the data will be stripped of any confidential information and stored in a private Github Repository.

In addition, the following materials are stored on a Google Drive and will be accessed by you during the study:

- PDFs of 2 research papers
- A video on how to complete one of the sections of the survey

Please be advised that this research study includes data storage in the U.S. As such, there is a possibility that information about you that is gathered for this research study may be accessed without your knowledge or consent by the U.S. government in compliance with the U.S. Patriot Act.

A.9 Ethics - Session 2 - Email to Participants Including Implied Consent Form

Hi <PARTICIPANT-NAME> !

Thank you for expressing interest in our study!

We are Cassandra Cupryk and Margaret-Anne Storey from the Computer Human Interaction and Software Engineering Lab (CHISEL) at the department of Computer Science at the University of Victoria, Canada as well as Paul Ralph from the department of Computer Science at the University of Dalhousie, Canada. The purpose of the study is to determine whether a tool that we've built can help Computer Science Master's and Ph.D. students better review research papers.

Thus, you are being invited to participate in the study entitled "Evaluation of the Empirical Standards Tool". We would appreciate the opportunity to gather data on how you'd use the Empirical Standards tool to review research papers as well as if you found the Empirical Standards tool to be beneficial.

We would be delighted if you would be willing to:

(A) let Us observe you for about 3 hours in an online group setting while you:
   a. use the tool to review 2 research papers
   b. download the 2 reviews (in form of text files) of the 2 research papers to your device
   c. upload the content of the 2 text files to a survey
   d. express your opinion of the tool in the survey, and
   e. discuss your opinions of the study with other people

(B) let Us analyze your reviews of the 2 research papers and your completed survey

(C) let Us record your video and your audio for the duration of the study

(D) answer questions about whether you found the tool useful for reviewing the 2 research papers

(E) answer questions about what improvements could be made to the tool

This is a purely academic research project with no commercial interests. As a master's student, I am required to conduct research as part of the requirements for a degree in Computer Science. It is being conducted under the supervision of Dr. Storey. You may

contact my supervisor at mstorey@uvic.ca if necessary.

Participants Selection

You are being asked to participate in this study because you have been identified as a Computer Science Master's or a Ph.D. student who is currently looking to improve how they review research papers. You have also indicated that you are willing to assist with our study by agreeing to being audio and video recorded.

What is involved

If you consent to voluntarily participate in this research:

- You will be asked to participate in an online study in a group setting with other individuals.
- You will be provided the website URL for the Empirical Standards tool.
- You will be asked to use the tool to review 2 research papers and download the 2 reviews to your device and upload the 2 reviews to the survey.
- You will be asked to complete a survey to express your opinion of the Empirical Standards tool.
- You will be asked to be a part of a discussion to express your opinion of the study.

Risks

There are no known or anticipated risks to you by participating in this research. Research participation is voluntary and you are under no obligation to participate. If at some point during the study you want to stop, you are free to do so without any negative consequences and any data collected up to that point will be discarded. In addition, choosing to participate or not to participate will not affect your grades, your standing, or your relationships, etc. The study will be completed in an online setting where other Computer Science Master's and Ph.D. students as well as study investigator(s) will be present.

Benefits

The potential benefits associated with the study include the development and improvement of a tool that can help Computer Science Master's and Ph.D. students better review research papers.

A.10 Ethics - HREB Application Approval - Session 1

University of Victoria | Research Administration Information System

Human Research Ethics Standard Application #21-0586

Summary

Study amendments summary

I. Instructions

NOTE: If your protocol is within 2 months of expiry, please delete this form and start the "Annual renewal with amendment" form instead, in case the protocol expires before the amendment is approved.

1. Complete the "Amendment summary and rationale" and "Unanticipated events" sections

2. Make changes to the application below

3. Mark changes to all amended appendices; delete old versions of appendices and upload new and amended appendices

4. Make sure you have the correct signatory/departmental head

5. Submit

II. Amendment summary and rationale

Provide a brief summary of all the changes you are proposing to make with a rationale why you need to make the proposed changes

We've changed the scope of the project. Previously, we were going to study whether the Empirical Standards tool (i.e. website) can help graduate students conduct and write about their research. Now, we are going to study whether the Empirical Standards tool can help Ph.D students better review other individuals' research papers.

Moreover, we need to make the following changes:

- Previously we were using semi-structured interviews that were essentially one-on-one meetings with participants for the study. Currently, we are using an online workshop, so there will be a group setting for the study.
- The workshop will still be audio and video recorded.
- The tool is the same as before.
- The online workshop will consist of: reviewing a paper using the tool and submitting that review of the paper, completing a survey evaluating the tool, and having a group discussion evaluating the tool.
- We are now planning on using an incentive(ex. gift card) to recruit students for the study.

III. Unanticipated events (TCPS 2 Article 6.15)

An unanticipated event includes any incidents, experiences, or outcomes that have not been previously accounted for in the approved protocol and which place participants, or others, at a greater risk (i.e. physical, psychological, economic, etc.) than was previously anticipated. An unanticipated event may have implications for the conduct of the study or the integrity of the research data.

Have there been any unanticipated events experienced with this research that have not previously been reported to HREB?

No

Application

A. Research team

1. Principal investigator (faculty, faculty supervising a student or post-doctoral researcher)

Principal Investigator is a faculty member, adjunct professor or sessional instructor. For more information please see the annotated guidelines.
If the project has more than one Principal Investigator (other than you) or more than one Principal Applicant, their names should be listed under section A.3 Research Team Members.

PI name

Margaret-Anne Storey

PI department
PI department. If more than one department, the department you are doing the research for.

Computer Science COSI

PI position
PI position at UVic

Faculty

2. Principal applicant (students & post-docs)

For further information about the distinction between the Principal Investigator and Principal Applicant, please see the annotated guidelines.
A Principal Applicant is an undergraduate student, graduate student or post-doctoral fellow who will be the lead researcher (for their thesis, dissertation, project, etc.) for this study. A Principal Applicant will be granted "View and edit" access by default, and will receive notifications related to the study. If the project has more than one Principal Applicant, the additional individuals should be listed under section A.3 Research Team Members.

Does this application have a principal applicant (UVic student or post-doc conducting this research for their academic degree)?

Yes

PA name

Cassandra Cupryk

PA email

cassandra.cupryk@gmail.com

PA department

Computer Science COSI

PA position

Master's student

PA phone

4632017203

PA graduate secretary's email (if the principal applicant is a graduate student. Leave blank otherwise.)

gradsec@csc.uvic.ca

Is the principal applicant conducting this research for their academic degree at UVic?

Yes

3. Research team members

Individuals and organizations involved in conducting your research. This includes co-principal investigators, additional principal applicants, co-investigators, other UVic students, assistants (paid or unpaid), community organizations, and clients. Team members

listed will have "no access" to application as a default. You cannot assign access to team members without Netlink ID. If they need a Netlink ID go to the Affiliate Identity Management System and click on the 'Sponsor' tab to start the process. Once you get the Netlink ID you have to re-enter their name and give access permission to the application.

List all current research team members (including any UVic students or research assistants who will use the received data or biological materials to fulfill UVic thesis, dissertation, or academic requirements) and assign level of access to the application. Inclusion here satisfies only UVic institutional requirements. If you grant "View and Edit" access to more than one person, be aware that the system will not notify users if and when others are making edits to the application.

DO NOT add the PI or PA to this table as that will cause technical permission issues.

Access: ✎ View and edit project 👁 View only ✉ Receive notifications $ Contribute funding

Name	Email	Role in the project	Institutional affiliation	✎	👁	✉	$
Cassandra Petrachenko	cpetrach@uvic.ca	Staff Collaborator: lab manager, helping with data analysis	Uvic	☑	☐	☐	☐
Alessandra Milani	amilani@uvic.ca	Student Collaborator: helping with research (theme coding, etc.)	UVic	☐	☐	☐	☐
Paul Ralph	paulralph@dal.ca	Collaborator	Dalhousie University	☐	☐	☐	☐
Enrique Larios	elarios.vargas@gmail.com	Professor Collaborator, help with research (planning the methodology, theme coding, etc.)	Uvic	☐	☐	☐	☐

B. Project information

1. Project title

Title for your research project. You may not submit two applications with the same title.

Evaluating A Software Engineering Empirical Standards Tool

2. Anticipated duration of the project

a. Anticipated start date for recruitment/data collection

The approximate start date to begin recruitment and data collection for your project should take into account the time it will take to complete and submit this application form and the period of four to six weeks required for ethical review. It is a violation of University of Victoria policy to begin recruitment and data collection before receiving HREB ethics approval.

Upon Approval

b. Anticipated end date for your research project

An approximate end date for recruitment and data collection.

December 2022

3. Is this application linked to one that has been recently submitted to the UVic Human Research Ethics Board?

No

4. Geographic location(s) of the study

Victoria, British Columbia

5. Keywords to categorize your research

Software Engineering | Empirical Research | | |

C. Project funding

1. Have you and/or research team members (their names must be listed under section A. Research team) applied for or been awarded funding for this project?

This information is used to permit the release of funds and to ensure proper reporting of research ethics approval to funding agencies. Please ensure the information in this table is correct.

No

2. Will this project receive funding from the US National Institute of Health (NIH)?

No

3. If you are a faculty member and have indicated above that you have applied for external funding, have you submitted a Research Application Summary Form to the Grants or Contracts unit in the Office of Research Services?

You must submit a research application summary form to the grants or contracts office every time you apply for external funding. Provide explanation, if you haven't done so.

Not applicable

Comments

D. Multi-jurisdictional research

1. Will this research be conducted under the auspices of another academic institution or health authority in BC (e.g. recruiting through their sites/departments/listservs/poster placement, etc.; involving staff, patients, health records; research team member affiliation)? The checklist may be useful, if unsure.

Research Ethics BC harmonized review (a single coordinated review with the other institution(s) listed) applies if A) you will be conducting research under the auspices of any of the institutions listed in REBC website (involving staff, patients, health records, sites and/or recruitment through their sites, including recruitment via poster placement), and/or B) when members of your research team consist of faculty, staff and students from any of the REBC institution(s).

No

2. Does the proposed research require Research Ethics Board (REB) approval from outside BC?

No

3. If this is a multi-jurisdictional research, please indicate your role in the research project (Check all that apply).

If you answered "Yes" to question D.1 please STOP completing this form and contact HRE office ethics@uvic.ca, 250-472-4321 or 250-472-4545 as soon as possible.

☐ Recruiting Participants

☐ Collecting data

☐ Analyzing data (with or without identifiers collected by you and/or your UVic research team members)

☐ Analyzing data that contain identifiers: data to be collected by non-UVic research team members as outlined in this application

☐ Analyzing data that does not contain identifiers: data to be collected by non-UVic research team members as outlined in this application

☐ Dissemination of results via publications, reports, conferences, internet, etc.

☐ Other

4. Additional information

There is no multi-jurisdictional research.

(Recruitment for the Second Option) Recruitment of registered Software Engineering Ph.D students will be done at the University of Victoria.

E. Other approvals and consultations

1. If additional request(s) for permission/approval are required please complete the section below (check all that apply)

Other approvals and consultations	Yes, approval uploaded	Yes, will provide as received	No approval required
a. School district, superintendent, principal, teacher			☑
b. Health authorities outside BC involving staff, patients, health records, sites and/or recruitment through their sites (including recruitment via poster placement)			☑
c. Other regional government authority			☑
d. Community group (e.g. formal organization, informal collective)			☑
e. UVic Biosafety Committee approval			☑
f. Other approval			☑

Please upload proof of having made request(s) for permission or any permission/approval documents that you received. Please forward approvals upon receiving them. Be assured that ethics approval may be granted prior to receipt of external approvals.

Comments

F. Scholarly review

1. What type of scholarly review has this research project undergone?

☐ External peer review (e.g. granting agency)

☑ Supervisory committee or supervisor - required for all student research projects

☐ None

☐ Other

G. Researcher(s) qualifications

1. In light of your research methods, the nature of the research, and the characteristics of the participants, what training, qualifications, or personal experiences do the principal investigator, the principal applicant, and/or your research team members have?
E.g. research methods course, language proficiency, committee experience, training on the equipment to be used.

Cassandra Cupryk, the principal applicant, has completed a variety of courses focusing on human computer interaction, research strategies, and study design. Her course work has prepared her to carry out the proposed study. This research will be conducted as a part of her Master's Project.

The project supervisor Dr. Margaret-Anne Storey and her collaborator Dr. Paul Ralph are both renowned for their outstanding work in computer science and software engineering research. They both have extensive experience supervising empirical studies.

2. Tri-Council Policy Statement - TCPS2 CORE Tutorial requirements
All UVic graduate students conducting research with human participants for their UVic project, thesis or dissertation are required to complete the Course on Research Ethics (CORE Tutorial) and provide evidence of ethics training by uploading a CORE completion certificate under this section.

List all current UVic graduate students (also listed under A.2 and A.3) involved in this research project for their UVic project, thesis or dissertation, and upload their Course on Research Ethics (CORE) tutorial certificate(s), if available. This CORE certification is required as of September 1, 2020 for new applications - see the human research ethics web page for more information.

Name	Email	Role in the project	CORE tutorial completion date
Cassandra Cupryk	ccupryk@uvic.ca	PA	June 5 2021

Supporting documents

tcps2_core_certificate.pdf *(Other approval, Name: TCPS 2: CORE, Version: Version 1)*; N 18, 2021

Comments

H. Research Involving the First Nations, Inuit and Métis Peoples of Canada

The TCPS2 (chapter 9) is designed to serve as a framework for the ethical conduct of research involving Aboriginal (including First Nations, Inuit and Métis) or Indigenous peoples, regardless of where they reside or whether or not their names appear on an official register. Its purpose is to ensure, to the extent possible, that research involving Indigenous peoples is premised on respectful relationships and encourages collaboration and engagement between researchers and participants.

This Policy acknowledges the role of the community in shaping the conduct of research that affects First Nations, Inuit, and Métis peoples. The nature and extent of community engagement should be determined through discussion with, and under the advisement of, the relevant community, taking into account relevant characteristics and protocols and the nature of the research.

The University of Victoria Indigenous Plan recognizes that research with Indigenous communities or involving Indigenous peoples must be conducted in a respectful and culturally appropriate manner, following protocols regarding entering community sites, engaging with communities, Elders and Knowledge Keepers, acknowledging cultural knowledge and cultural property, and disseminating research findings.

1. Conditions of the research

a. Will you be conducting research that is situated on any of the following kinds of lands or waterways: First Nation reserves, Indigenous settlements, Indigenous lands under self-government agreements, territories with Indigenous land claims agreements, or other lands designated by Federal, Provincial, or local governments as Indigenous territory?

No

b. Do any of the criteria for participation include belonging to an Indigenous nation, community, group of communities, or organization, including urban Indigenous populations?

No

c. Does the research seek input from participants regarding Indigenous cultural heritage, cultural practices, artifacts, Indigenous or traditional knowledges, or distinct characteristics of Indigenous experience or reality?

No

d. Will Indigenous identity or membership in an Indigenous community or group (e.g. Métis Nation) be used as a variable for the purposes of analysis?

No

e. Will the results of the research make specific reference to Indigenous communities, homelands and/or waterways, peoples, languages, histories or cultures?

No

2. Indigenous engagement

a. Processes and protocols for engagement differ across communities, organizations, committees, and groups, as well as across different research contexts. Describe the process that you have followed with respect to Indigenous engagement.
Include any documentation of collaboration (e.g. formal research agreement, letter of approval, email communications, advisory committee, mentorship, etc.) and the role or position of those consulted (e.g. Elder, Knowledge Holder, governing body, Chief, etc.), including their names, if appropriate.

Not applicable.

b. Explain how Indigenous community members will be meaningfully involved throughout the research process, from research design to knowledge sharing.
Outline the plan, as developed with the community, for the outcomes of the research, including research data ownership, sharing, storage, and governance.

Not applicable.

c. If you have answered "yes" to any of the questions in H.1 but have not yet engaged with the community, committee, organization, or group, please explain why not and outline how you plan to conduct a study that respects Indigenous communities and participants in the absence of prior engagement.

3. Comments

I. International research

1. Will this study be conducted in a country other than Canada?

No

J. Description of research project

1. Briefly describe in non-technical language

a. The research objective(s) and question(s)

We aim to understand whether the Empirical Standards Tool can help Ph.D students better review research papers.

b. The importance and contributions of the research

The Empirical Standards tool is a website composed of multiple online checklists with essential information that would provide guidance to Software Engineering Ph.D students when they are reviewing research papers.

The major contribution of this research is to discover whether the Empirical standards tool can help Software Engineering Ph.D students review research papers. This information can then be used to improve the Empirical Standards tool so that the tool can better serve Software Engineering Ph.D students.

c. If applicable, provide background information or details that will enable the Research Ethics Board to understand the context of the study when reviewing the application

This research will be conducted as part of a Master's Project of the principal applicant at the University of Victoria. The research information will be accessed only by the researchers.

K. Recruitment

1. Participant details

Provide details of your participants

a. Briefly describe the target population(s) for recruitment
Ensure that all participant groups are identified (e.g. group 1 - teacher, group 2 - administrators, group 3 - parents).

Software Engineering Ph.D students.

b. Why is each population or group of interest?

The tool is aimed to primarily help Software Engineering Ph.D students who are currently looking to improve how they review research papers.

c. What are the salient characteristics of the participants for your study (e.g. age, gender, ethnicity, class, position, etc.)?
List all inclusion and exclusion criteria you are using.

Software Engineering Ph.D students who are currently looking to improve how they review research papers.

d. What is the desired number of participants for each group?

6-12 participants

2. Recruitment and process

Provide details of your recruitment process

a. List all source for information used to contact potential participants
E.g. personal contacts, listserves, publicly available contact information, etc. Clarify which sources will be used for which participant groups.

Registered Software Engineering Ph.D students in Software Engineering classes at University of Victoria. We will use official Uvic email lists as well as some Uvic social media channels (ex. Reddit, Facebook, Discord) to approach students.

b. List all methods of recruitment
E.g. in-person, by telephone, letter, snowball sampling, word-of-mouth, advertisement, etc. If you will be using "snowball" sampling, clarify how this will proceed (i.e. will participants be asked to pass on your study information to other potential participants?). Clarify which methods will be used for which participant groups.

Firstly, we will be contacting moderators of a listserv or a social media channel in order to recruit participants as explained below:
1. An email will be sent to a moderator of a listserv (Appendix_09_Email_To_Moderator_Listserv) in order to inquire if an email can be sent out to the students on the moderator's listserv about information of the study. If the moderator agrees to allow us to send an email with the information of the study (Appendix 07 – Posting of Study) to the emails on the listserv, then the email will be sent out by the moderator. The researchers will not have any access to any contact information or the emails on the listserv. The moderators are the only individuals who will have access to the potential participants' contact information and who will reach out to the potential participants.
2. An email or message will be sent to a moderator of a social media channel (Appendix_08_Email_To_Moderator_Social_Media) in order to inquire if the information of the study can be posted in the channel. If the moderator agrees to allow us to post information of the study in the channel, then the information of the study (Appendix 07 – Posting of Study) will be posted in the channel by the moderator. The researchers will not have access to any contact information of the users in the channel. The moderators are the only individuals who will have access to the potential participants' contact information and who will reach out to the potential participants.

Any students that are interested in the study will send us an email. We will then send an email (Appendix 01 – Email to the Participants) to the participants who have reached out to us. In the email, we will ask the participants to complete the signed consent form (Appendix 03 - Signed Consent Form) if they are interested in the study. The participant will only be allowed to participate in the study if they've signed the consent form.

c. If you will be using personal and/or private contact information to contact potential paticipants (as stated above), have the potential participants given permission for this, or will you use a neutral third party to assist you with recruitment?
Note that this is not a concern when public and/or business contact information is used.

We will rely on neutral third parties to interact with the Ph.D students. The moderators will not grant us access to the potential participants' contact information. The recruitment will only proceed through the support of moderators.

d. Who will recruit/contact participants?
E.g. researcher, assistant, third party, etc. Clarify this for each participant group.

The principal investigator, the principal applicant, and co-investigators will contact the moderators of the listserv or social media channel. The moderators will then send out an email containing information of the study or post information of the study in their social media channel. In the posting of the study, interested participants are asked to contact the researcher via an email included in the posting of the study.

e. List and explain any relationship between the members of the research team (including third party recruiters or sponsors/clients of the research) and the participant(s) (e.g. acquaintances, colleagues)
Complete section 3 (Power relationship) if there is potential for a power relationship or a perceived power relationship (e.g. instructor-student, manager-employee, etc.). If you have a close relationship with potential participants (e.g. family member, friend, close colleague, etc.) clarify the safeguards that you will put in place to mitigate any potential pressure to participate.

The investigators have no known relationship with any of the participants, but may be supervisors of the students or teaching them courses.

f. In chronological order (if possible) describe the steps in the recruitment process
Include how you will screen potential participants, where applicable. Consider where in the process permission of other bodies may be required.

Firstly, we will be contacting moderators of a listserv or a social media channel in order to recruit participants as explained below:
1. An email will be sent to a moderator of a listserv (Appendix_09_Email_To_Moderator_Listserv) in order to inquire if an email can be sent out to the students on the moderator's listserv about information of the study. If the moderator agrees to allow us to send an email with the information of the study (Appendix 07 – Posting of Study) to the emails on the listserv, then the email will be sent out by the moderator. The researchers will not have any access to any contact information or the emails on the listserv. The moderators are the only individuals who will have access to the potential participants' contact information and who will reach out to the potential participants.
2. An email or message will be sent to a moderator of a social media channel (Appendix_08_Email_To_Moderator_Social_Media)

in order to inquire if the information of the study can be posted in the channel. If the moderator agrees to allow us to post information of the study in the channel, then the information of the study (Appendix 07 – Posting of Study) will be posted in the channel by the moderator. The researchers will not have access to any contact information of the users in the channel. The moderators are the only individuals who will have access to the potential participants' contact information and who will reach out to the potential participants.

Any students that are interested in the study will send us an email. We will then send an email (Appendix 01 – Email to the Participants) to the participants who have reached out to us. In the email, we will ask the participants to complete the signed consent form (Appendix 03 - Signed Consent Form) if they are interested in the study. The participant will only be allowed to participate in the study if they've signed the consent form.

Please upload all the supporting documents relevant to the recruitment methods identified in this section
Examples of supporting documents: email recruitment script, poster, invitation letter, etc. Where draft versions are uploaded please ensure that final versions are submitted when available. If final versions differ significantly after you have obtained research ethics approval, you will need to submit a Request for Amendment.

Supporting documents

Appendix_01_Email_Invitation_To_Participants_CC_v2.pdf
(Recruitment document, Name: A01 - Email Invitation to Participants, Version: V2); M 11, 2022

Appendix_08_Email_To_Moderator_Social_Media_CC_v2.pdf
(Recruitment document, Name: A08 - Email to Moderator Social Media, Version: V2); M 11, 2022

Appendix_09_Email_To_Moderator_Listserv_CC_v2.pdf
(Recruitment document, Name: A09 - Email to Moderator Listserv, Version: V2); M 11, 2022

Appendix_07_Posting_of_Study_CC_v2.pdf *(Recruitment document, Name: A07 - Posting of Study, Version: V2)*; M 11, 2022

3. Power relationship (dual-role and power-over)

If you are completing this section, please refer to the guidelines for ethics in dual-role research for teachers and other practitioners and the TCPS2, article 3.1 and article 7.4.

Are you or any of your co-researchers in any way in a power relationship, including dual-roles, that could influence the voluntariness of a participant's consent? Could you or any of your co-researchers potentially be perceived to be in a power relationship by potential participants?
Examples of "power relationships" include teachers-students, therapists-clients, supervisors-employees and possibly researcher-relative or researcher-close-friend where elements of trust or dependency could result in undue influence.

No

L. Data collection methods

1. Data collection methods

Use the following sections in ways best suited to explain your project. If you have more than one participant group, be sure to explain which participant group(s) will be involved in which activity/activities or method(s).
If this research will/may include in-person activities during the global pandemic, you must fulfill the requirements supporting in-person research with human participants. Please complete relevant section of the application and appendices with the information outlined in the current UVic Human Research Ethics COVID-19 Bulletin, under the human research ethics webpage.

a. Which of the following methods will be used to collect data? Check all that apply

If this research will/may include in-person activities during the global pandemic, you must fulfill the requirements supporting in-person research with human participants. Please complete relevant section of the application and appendices with the information outlined in the current UVic Human Research Ethics COVID-19 Bulletin, under the human research ethics webpage.

- ☑ i) Interviewing participants
  - ☐ In person
  - ☐ By telephone
  - ☑ Conducting group interviews or discussions (including focus group)

☑ Using web-based technology

Explain and provide the name of the web-based technology or technologies (e.g., Skype, Bluejeans, etc.). For more information on platforms, programs, security, etc. when conducting research with participants virtually see the UVic FAQ and the U.S. Freedom Act advisory below.
If using a web program (online surveys, video conferencing etc.) with a server located in the United States (e.g. SurveyMonkey), or if there are other reasons that the data will be stored in the US (e.g. use of US-based cloud technology, sharing data with US colleagues, etc.), you must inform participants that their responses may be accessed via the U.S. Freedom Act. Please add the following to the consent form(s): "Please be advised that this research study includes data storage in U.S.A. As such, there is a possibility that information about you that is gathered for this research study may be accessed without your knowledge or consent by the U.S. government, in compliance with the U.S. Freedom Act."

Audio and video data will recorded via Zoom.
A facilitated discussion will be done during the study and may use software such as Miro, etc.

If this research will/may include in-person activities during the global pandemic, you must fulfill the requirements supporting in-person research with human participants. Please complete relevant section of the application and appendices with the information outlined in the current UVic Human Research Ethics COVID-19 Bulletin, under the human research ethics webpage.

☑ ii) Administering a questionnaire or survey

☐ In person

☐ By telephone

☐ Email

☐ Mail back

☑ Web-based

Explain and provide the name of the web-based technology or technologies (e.g., SurveyMonkey), and see the U.S. Freedom Act advisory below.
If using a web program (online surveys, video conferencing etc.) with a server located in the United States (e.g. SurveyMonkey), or if there are other reasons that the data will be stored in the US (e.g. use of US-based cloud technology, sharing data with US colleagues, etc.),you must inform participants that their responses may be accessed via the U.S. Freedom Act. Please add the following to the consent form(s): "Please be advised that this research study includes data storage in U.S.A. As such, there is a possibility that information about you that is gathered for this research study may be accessed without your knowledge or consent by the U.S. government, in compliance with the U.S. Freedom Act."

Survey will administered using SurveyMonkey.

☐ Other

☑ iii) Administering a computerized task (describe in section L.1b and/or upload documents)

☑ iv) Observing participants. In section L.1b describe who and what will be observed. Include where observations will take place. If applicable, upload an observational collection sheet for review.

☑ v) Recording of participants and data

☑ Audio

☑ Video

How are the video images going to be used?

☑ Images used for analysis

☐ Images used in disseminating results (include release to use participant images' in consent materials)

☐ Photos or slides

☑ Note taking

☐ Flipcharts

☐ Data collection sheets (upload)

☐ Other

Refers to information/data that was originally gathered for a purpose other than the proposed research and is now being considered for use in research (e.g. patient or school records, personal writings, lesson plans, etc.).

☐ vi) Using human samples (e.g. saliva, urine, blood, hair)

☐ vii) Using specialized equipment/machines (e.g. ultrasound, EEG, prototypes, etc.) or other (e.g. testing instruments that are not surveys or questionnaires)

☐ viii) Using other testing equipment not captured under other categories

E.g. artifacts, paintings, drawings, photos, slides, art, journals, writings, etc.

☑ ix) Collecting materials supplied by, or produced by, the participants

Please specify

- We will be asking the students to write a review of the research paper we provide to them and we will collect all these reviews.
- We'll also be collecting the surveys completed by the students.

Refers to information/data that was originally gathered for a purpose other than the proposed research and is now being considered for use in research (e.g. patient or school records, personal writings, lesson plans, etc.).

☐ x) Analyzing secondary data or secondary use of data

☐ xi) Other

b. Provide a sequential description of the procedures/methods to be used in your research study
Be sure to provide details for all methods checked in section L.1. Clarify which procedures/methods will be used for each participant group. Indicate which methods, if any, will be conducted in a group setting. List all of the research instruments and interview/focus group questions, and append copies (if possible) or detailed descriptions of all instruments. If not yet finalized, provide drafts or sample items/questions..

If using a web program (online surveys, video conferencing etc.) with a server located in the United States (e.g. SurveyMonkey), or if there are other reasons that the data will be stored in the US (e.g. use of US-based cloud technology, sharing data with US colleagues, etc.),you must inform participants that their responses may be accessed via the U.S. Freedom Act. Please add the following to the consent form(s): "Please be advised that this research study includes data storage in U.S.A. As such, there is a possibility that information about you that is gathered for this research study may be accessed without your knowledge or consent by the U.S. government, in compliance with the U.S. Freedom Act."

Firstly, prior to the study, we will ask the participants to complete the consent form (Appendix 03 - Signed Consent Form). Once the participant signs and returns the consent form, they will attend the online study at the online location (i.e. University of Victoria Zoom).

Once the participants arrive to the study, we will follow the general script for the study (Appendix 02 – Overview of Workshop Script).

The participants will be reminded that they will be audio recorded for the duration of the study. The participants will also be reminded that they need to share their screen and that their screen will be recorded for the duration of the study. Participants will be reminded that they will be required to remove any personal identifying information (e.g., close email applications or websites, turn off chat applications, etc.) from their computer screen.

1. For the first part of the study, the participants will be asked to use the Empirical Standards tool to help them review a research paper. The link to the Empirical Standards tool will be provided. The participant will then submit the review of the research paper.

2. For the second part of the study, the participants will be asked to complete a survey (Appendix 04 – Draft of Survey) hosted on SurveyMonkey in order to share their opinion of the Empirical Standards tool.

3. For the last part of the study, the participants will participate in a facilitated discussion where all the participants will express their opinions of the Empirical Standards tool.

The principal applicant will be online over the whole duration of the study. The audio from the study will be recorded and later-on transcribed into text. The principal applicant will take digital notes during the study. The audio, surveys, and the reviews will be stored on the principal applicant's Uvic OneDrive Account.

c. Where will participation take place for each data collection method/procedure?
Provide specific location (e.g. UVic classroom, private residence, participant's workplace). Clarify the locations for each participant group and/or each data collection method.

Participant's workplace, home, or a place where they are comfortable attending the online workshop.

d. For each method, and in total, how much time will be required of participants?

Clarify this for each participant group, each data collection method, and any other research related activities.

3 hours in total:
- Participants reviewing a research paper using the tool and submitting that review of the research paper.
- Participants completing a survey evaluating the tool.
- Participants having a group discussion evaluating the tool.

e. Will participation take place during participants' office work/hours or instructional time?

No

2. Data collection materials checklist

Data collection methods checklist

- ☑ Standardized instrument
- ☑ Survey
- ☐ Questionnaire
- ☑ Interview and/or focus group questions
- ☑ Observation protocols
- ☐ Other

Please make sure that you have uploaded all the documents relevant to this section. Add any other documents that you think may be relevant to this section.
Where draft versions are appended please ensure that final versions are submitted when available. If final versions differ significantly after you have obtained research ethics approval, you will need to submit a Request for Modification.

Supporting documents

Appendix_06_Description of the Tool and Instructions To Use The Tool_CC_v2.pdf
(Data collection instrument, Name: A06- Description of Tool and Instruction, Version: V2); M 11, 2022

Appendix_05_Protocol_CC_v2.pdf *(Data collection instrument, Name: A05 - Protocol of Study, Version: V2)*; M 11, 2022

Appendix_04_Draft_Survey_CC_v2.pdf *(Data collection instrument, Name: A04 - Draft of Survey, Version: V2)*; M 11, 2022

Appendix_02_Overview_of_Workshop_Script_CC_v2.pdf
(Data collection instrument, Name: A02 - Overview of Workshop Script, Version: V2); M 11, 2022

M. Possible benefits, inconveniences, and risks of harm to participants

1. Benefits

Identify any potential or known benefits associated with participation and explain below
Keep in mind that the anticipated benefits should outweigh any potential risks.

- ☐ To the participants
- ☐ To society
- ☑ To the state of knowledge

Please explain

Will provide a deep understanding of whether the tool can help Software Engineering Ph.D students with reviewing other individuals' research papers.

2. Inconveniences

Identify and describe any known or potential inconveniences to participants

Consider all potential inconveniences, including total time devoted to the research.

No inconveniences apart from time commitment of maximum 3 hrs.

3. Level of risk

The TCPS 2 article 6.12 definition of "minimal risk research" is as follows: 'Research in which the probability and magnitude of possible harms implied by participation in the research is no greater than those encountered by the participant in those aspects of their everyday life that relate to the research.'

Based on this definition, do you believe your research qualifies as 'minimal risk research'?

Yes

Explain your answer with reference to the risks of the study and the vulnerability of the participants

Participants are being asked to volunteer around 3 hours of their time.

4. Estimate of risks of harm

Potential risks of harm	Very unlikely	Possibly	Likely
a. Emotional or psychological discomfort, such as feeling demeaned or embarrassed due to the research	☑		
b. Fatigue or stress	☑		
c. Social risks, such as stigmatization, loss of status, privacy and/or reputation	☑		
d. Physical risk such as falls	☑		
e. Economic risks (e.g. job security, salary loss, etc.)	☑		
f. Risk of incidental findings (see article 3.4 of the TCPS 2 for more information)	☑		
g. Other risks	☑		

Consider the inherent foreseeable risks associated with your research protocol and complete the table below by selecting the options that best fit the potential risks listed below. Be sure to take into account the vulnerability of your target population(s) if applicable.

If other risks, please specify

5. Possible risks of harm

If you indicated in item 4 (a) to (g) that any risks of harm are possible or likely, please explain below

a. What are the risks?
I.e. elaborate on risks you have identified above.

b. What will you do to try to minimize, mitigate, or prevent the risks?

c. How will you respond if the harm occurs?
I.e. what is your plan?

d. If you have indicated that there is a risk of incidental findings in item 4 (f), please outline your proposed protocol for information and /or action

e. If one of your participant groups could be considered vulnerable, please describe any specific considerations you have built into the protocol to address this

6. Risk to researcher(s)

Does this research study pose any risks to the researchers, assistants and data collectors?

No

7. Deception

Will participants be fully informed of everything that will be required of them prior to the start of the researcher session?

Yes

If not, complete the Request to use Deception form on the ORS website

N. Incentives, reimbursement and compensation

1.Is there any incentive, monetary or otherwise, being offered for participation in the research (e.g. gifts, honorarium, course credits, etc.)?

Yes

Explain the nature of each incentive and why you consider it necessary
Also consider whether the amount or nature of the incentive could be considered a form of undue inducement or affect the voluntariness of consent. Clarify which participant groups will be provided with which incentives.

$20 Amazon Gift Cards may be offered to this participant group in order to attract more participants to the study.

2. Is there any reimbursement or compensation for participating in the research (e.g. for transportation, parking, childcare, etc.)?

No

3. Explain what will happen to the incentives, reimbursement or compensation if participants withdraw during data collection or any time thereafter
E.g. compensation will be pro-rated, full compensation will be given, etc.

If the participant withdraws from the study they will get 100% of the incentive. This will ensure that participants are not forced or will not feel forced or coerced in any way to complete the study.

O. Free and informed consent

Consent encompasses a process that begins with initial contact and continues through to the end of the research process.
Consult article 3.2 of the TCPS 2 and appendix V of the guidelines for further information.

1. Participant's capacity (competence) to provide free and informed consent

Capacity refers to the ability of prospective or actual participants to understand relevant information presented about a research project, and to appreciate the potential consequences of their decision to participate or not participate. See the TCPS 2, chapter 3, section C, for further information.

Identify your potential participants (check all that apply)

a. Competent

- ☑ i) Competent adults
- ☐ ii) A protected or vulnerable population (e.g. inmates, patients)
- ☐ iii) Competent youth aged 13 to 18

☐ iv) Competent children under 13 (who are able to provide fully informed consent)

b. Non-competent

☐ i) Non-competent adults

☐ ii) Non-competent youth

☐ iii) Non-competent children (young children and/or children with limited abilities to provide fully informed consent)

2. Means of obtaining and documenting consent and/or assent:

Check all that apply
When completing this section make sure that you consider all of your participant groups, upload copies of relevant materials and complete section O3.

☑ Signed consent

Upload consent form(s) in section O.5 or section S - see template.

☐ Verbal consent

☐ Letter of information for implied consent (e.g. anonymous, mail back or web-based survey)

☐ Signed or verbal assent for non-competent participants

☐ Other means

☐ Consent will not be obtained

☐ Signed consent from the parents/guardians for youth/child participants

☐ Information letters for the parents/guardians of youth/child participants

3. Informed consent

Describe the exact steps (chronological order) that you will follow in the process of explaining, obtaining, and documenting informed consent
Ensure that consent procedures for all participant groups are identified (e.g. group 1 - teachers, group 2 - parents, group 3 - students). Be sure to indicate when participants will first be provided with the consent materials (e.g. prior to first meeting with the researcher?). If consent will not be obtained, explain why not with reference to the TCPS 2 articles 3.5 and 3.7.

We will send an Invitation to Moderator of a Social Media channel (APPENDIX 08 - Email to Moderator - Social Media) to get permission to post about the study in their channel or we will send an Invitation to Moderator of a Listserv (APPENDIX 09 - Email to Moderator - Listserv) to ask about sending an email about the study to all students (as an example: engn-announce-bounces@lists.uvic.ca). With the moderator's permission we will send out or post information about the study (APPENDIX 07 - Posting of the Study). Interested parties will be asked to contact us via email. Potential participants will be sent an Invitation to Participate (APPENDIX 01 - Email Invitation to Participants) via email with a Signed Consent Form (APPENDIX 03 - Signed Consent Form) attached. The participants who choose to participate will be asked to complete the Signed Consent Form and send the signed consent form back. The participant will only be allowed to participate in the study if they've signed the consent form.

4. Ongoing consent

Will your research occur over multiple occasions or an extended period of time (including review of transcripts)?

No

5. Participant's right to withdraw

Article 3.1 of the TCPS 2 states that participants have the right to withdraw at any time and can withdraw their data and human biological materials.

a. Describe what participants will be told about their right to withdraw from the research at any time (i.e., who to contact and how)
If compensation is involved, explain what participants will be told about compensation if they withdraw. If you have different participant groups and/or different data collection methods, clarify the different procedures for withdrawing as necessary.

Participants will be told through the signed consent form (APPENDIX 03 - Signed Consent Form) that they may stop participating in the study at any point without explanation.

b. What will happen to a person's data if they withdraw part way through the study or after the data have been collected/submitted?
If applicable, include information about visual data such as photos or videos. If you have different participant groups and/or different data collection methods, clarify the different procedures for withdrawing as necessary. Ensure this information is included in the consent documents.

☐ Participant will be asked if they agree to the use of their data

☑ It will not be used in the analysis and will be destroyed

☐ It is logistically impossible to remove individual participant data (e.g. anonymously submitted data)

☐ When linked to group data (e.g. focus group discussions), it will be used in summarized form with no identifying information

Please make sure that you have uploaded all the documents relevant to this section. Add any other documents that you think may be relevant to this section.
Where draft versions are appended please ensure that final versions are submitted when available. If final versions differ significantly after you have obtained research ethics approval, you will need to submit a Request for Modification.

Supporting documents

Appendix_03_Signed_Consent_Form_CC_v2.pdf
(Consent/assent form, Name: A03 - Signed Consent Form, Version: V2); M 11, 2022

P. Anonymity and confidentiality

1. Anonymity

Anonymity means that no one, including the principal investigator, is able to associate responses or other data with individual participants.

a. Will the participants be anonymous in the data gathering phase of research?

No

b. Will the participants be anonymous in the dissemination of results (be sure to consider use of video, photos)?

Yes

2. Confidentiality

Confidentiality means the protection of the person's identity (anonymity) and the protection, access, control and security of their data and personal information during the recruitment, data collection, reporting of findings, dissemination of data (if relevant) and after the study is completed (e.g. storage). The ethical duty of confidentiality refers to the obligation of an individual or organization to safeguard entrusted information. The ethical duty of confidentiality includes obligations to protect information from unauthorized access, use, disclosure, modification, loss or theft.

a. Are there any limits to protecting the confidentiality of participants?

Yes, there are some limits to the researcher's ability to protect the confidentiality of participants (check all that apply)

E.g. focus groups. The researcher cannot guarantee confidentiality.

☑ Limits due to the nature of group activities

The nature or size of the sample from which participants are drawn makes it possible to identify individual participants (e.g. school principals in a small town, position within an organization).

☐ Limits due to context

The procedures for recruiting or selecting participants may compromise the confidentiality of participants (e.g. participants are identified or referred to the study by a person outside the research team).

☐ Limits due to selection

E.g. legal or professional.

☐ Limits due to legal requirements for reporting

E.g. when there will be data storage in the United States. When using USA based data instruments and data storage systems researchers are responsible for determining if this applies.

☑ Limits due to local legislation such as the U.S. Freedom Act

☐ Other

b. If confidentiality will be protected, describe the procedures to be used to ensure the anonymity of participants and for preserving the confidentiality of their data (e.g. pseudonyms, changing identifying information and features, coding sheet, etc.)
If you will use different procedures for different participant groups and/or different data methods be sure to clarify each procedure.

The workshop is done within a group setting, so the researcher cannot guarantee confidentiality.

However, after the study, the participant's data will be access restricted, password protected, anonymized and will be kept confidential. The researchers are the only ones who will have access to this data. Any identifying information will be removed or generalized so that it is not possible to link the information back to the participant before the public dissemination of results.

c. If there are limits to confidentiality indicated in section P.2.a, explain what the limits are and how you will address them with the participants
If there are different procedures for different participant groups and/or different data collection methods, be sure to clarify each procedure.

The workshop is done within a group setting, so the researcher cannot guarantee confidentiality. The participants will be made aware that this study is a group setting in the signed consent form, thus, they will know that their confidentiality cannot be guaranteed.

In addition, this research study includes data storage in the U.S. As such, there is a possibility that information about the participant that is gathered for this research study may be accessed without the participant's knowledge or consent by the U.S. government in compliance with the U.S. Patriot Act. The participant will be made aware of this information in the signed consent form.

Q. Data management

1. Use(s) of data

a. What use(s) will be made of all types of data collected (field notes, photos, videos, audiotapes, transcripts, etc.)?

The data will be used to identify the problems that arise when Software Engineering Ph.D students use the Empirical Standards tool. The problems will be noted in order to make improvements to the tool. The data will be anonymized, analyzed, and then used to describe study findings. As part of our commitment to open science, study instruments and anonymized results will be available online for replication purposes.

b. Will your research data be analyzed, now or in future, by yourself for purposes other than this research project?

No

c. Will your research data be analyzed, now or in future, by other persons for purposes other than explained in this application?

No

2. Commercial purposes

Do you anticipate that this research will be used for a commercial purpose?

No

3. Maintenance and disposal of data

Describe your plans for protecting data during the project, and for preserving, archiving, or destroying all the types of data associated with the research (e.g. paper records, audio or visual recordings, electronic recordings, coded data) after the research is completed:

a. Means of storing and securing data
E.g. encryption, password protected computer files, locked cabinet, separation of key codes from raw data etc.

The participant's confidentiality and the confidentiality of their data will be access restricted and password protected. The participant's information will be anonymized and will be kept confidential. The researchers are the only ones who will have access to this data.

All of the data will be stored on a University of Victoria Microsoft Office 365 OneDrive Account. OneDrive is built on the Microsoft Office 365 hyper-scale, enterprise-grade cloud, which delivers advanced security and compliance capabilities. The data is encrypted in transit and at rest. All OneDrive files in the University of Victoria Microsoft 365 account are stored in Canada, however, some automated processing may occur outside of Canada. After conducting the interviews, the data will be stripped of any confidential information and stored in a private Github Repository. All notes will be taken digitally and stored in the Uvic OneDrive account as well.

b. Location of storing data

Include location of data-storage servers if using web-based technology.

Electronic Information: All OneDrive files in the University of Victoria Microsoft 365 account are stored in Canada, however, some automated processing may occur outside of Canada. Email correspondence will be saved on the University of Victoria email servers. Data stored on Github may have servers located U.S.A. and may be accessed under the US Patriot Act.

c. Duration of data storage

If data will be kept indefinitely, explain why this is necessary and state whether the data will contain identifiers or links to identifiers.

Two years.

d. Methods of destroying or archiving data

If archiving data, please describe measures to secure or protect the data. If the archiving will involve a third party (e.g. library, community agency, Aboriginal band, etc.) please provide details.

Electronic Information: Data and backups will be deleted from the repository.

4. Dissemination

How do you anticipate disseminating the research results? (check all that apply)

- ☑ Thesis/dissertation/class presentation
- ☑ Presentations at scholarly meetings
- ☑ Internet (students: most UVic theses are posted on 'UVicSpace' and can be accessed by the public)
- ☐ Media (e.g. newspaper, radio, TV)
- ☐ Directly to participants and/or groups involved
- ☑ Published article, chapter or book
- ☐ Other

R. Conflict of interest

1. Apart from a declared dual-role relationship (section K.3), are you or any of the research team members in a perceived, actual or potential conflict of interest regarding this research project (e.g. partners in research, private interests in companies or other entities)?

No

S. List of uploaded documents

Review the document requirements list and the uploaded documents to ensure that you have all the applicable documents. Make sure to remove all duplicates. Upload appendices as individual documents, instead of clustering appendices under one attachments. Incomplete applications and applications with incorrectly uploaded appendices will not be reviewed. You will be notified in this case.

App. version	Section	Descriptive name	File name	Type of document	Date uploaded	File versi
V1.2	G.	TCPS 2: CORE	tcps2_core_certificate.pdf	Other approval	Nov 18, 2021 7:17:48 AM	Versi 1
V1.2	K.	A01 - Email Invitation to Participants	Appendix_01_Email_Invitation_To_Participants_CC_v2.pdf	Recruitment document	Mar 11, 2022 2:14:19 PM	V2

V1.2	K.	A08 - Email to Moderator Social Media	Appendix_08_Email_To_Moderator_Social_Media_CC_v2.pdf	Recruitment document	Mar 11, 2022 2:15:12 PM	V2
V1.2	K.	A09 - Email to Moderator Listserv	Appendix_09_Email_To_Moderator_Listserv_CC_v2.pdf	Recruitment document	Mar 11, 2022 2:15:42 PM	V2
V1.2	K.	A07 - Posting of Study	Appendix_07_Posting_of_Study_CC_v2.pdf	Recruitment document	Mar 11, 2022 2:16:23 PM	V2
V1.2	L.	A06- Description of Tool and Instruction	Appendix_06_Description of the Tool and Instructions To Use The Tool_CC_v2.pdf	Data collection instrument	Mar 11, 2022 2:18:08 PM	V2
V1.2	L.	A05 - Protocol of Study	Appendix_05_Protocol_CC_v2.pdf	Data collection instrument	Mar 11, 2022 2:18:37 PM	V2
V1.2	L.	A04 - Draft of Survey	Appendix_04_Draft_Survey_CC_v2.pdf	Data collection instrument	Mar 11, 2022 2:19:01 PM	V2
V1.2	L.	A02 - Overview of Workshop Script	Appendix_02_Overview_of_Workshop_Script_CC_v2.pdf	Data collection instrument	Mar 11, 2022 2:19:31 PM	V2
V1.2	O.	A03 - Signed Consent Form	Appendix_03_Signed_Consent_Form_CC_v2.pdf	Consent /assent form	Mar 11, 2022 2:20:29 PM	V2

T. Signatory/Departmental sign-off

Select the Chair/Director/Dean or their designate to sign-off on this application for submission. Once signed-off, the application will be submitted to the Human Research Ethics Board for review.

By signing-off the application, the signatory is affirming that adequate research infrastructure is available for the conduct and completion of this research project.

Signatory name

Sudhakar Ganti

A.11 Ethics - HREB Application Approval - Session 2

University of Victoria | Research Administration Information System

Human Research Ethics Standard Application #21-0586

Summary

Study amendments summary

I. Instructions

NOTE: If your protocol is within 2 months of expiry, please delete this form and start the "Annual renewal with amendment" form instead, in case the protocol expires before the amendment is approved.

1. Complete the "Amendment summary and rationale" and "Unanticipated events" sections

2. Make changes to the application below

3. Mark changes to all amended appendices; delete old versions of appendices and upload new and amended appendices

4. Make sure you have the correct signatory/departmental head

5. Submit

II. Amendment summary and rationale

Provide a brief summary of all the changes you are proposing to make with a rationale why you need to make the proposed changes

NEW UPDATES:
1. Changing how the participants will be recruited. They will now be recruited via Twitter.
2. Changed from signed consent form to obtaining implied consent. The reason being is that physically signing a physical consent form may be difficult for Ph.D. and master's students to achieve without a printer or tablet. A physical signature on a PDF creates a possible obstacle for students.
3. Changed incentive from a $20 Amazon gift card to a $60 Amazon Gift Card.
4. Changed duration of data storage from 2 years to 5 years.

PREVIOUS UPDATES:
Some of the materials accessed by the participants during the study will be stored on Google Drive.
More in-depth summary of how the participants will be recruited .
More in-depth summary of the materials used during the study(ex. pdfs of: the presentation, the survey, etc.).

III. Unanticipated events (TCPS 2 Article 6.15)

An unanticipated event includes any incidents, experiences, or outcomes that have not been previously accounted for in the approved protocol and which place participants, or others, at a greater risk (i.e. physical, psychological, economic, etc.) than was previously anticipated. An unanticipated event may have implications for the conduct of the study or the integrity of the research data.

Have there been any unanticipated events experienced with this research that have not previously been reported to HREB?

No

Application

A. Research team

1. Principal investigator (faculty, faculty supervising a student or post-doctoral researcher)

Principal Investigator is a faculty member, adjunct professor or sessional instructor. For more information please see the annotated guidelines.

If the project has more than one Principal Investigator (other than you) or more than one Principal Applicant, their names should be listed under section A.3 Research Team Members.

PI name

Margaret-Anne Storey

PI department

PI department. If more than one department, the department you are doing the research for.

Computer Science COSI

PI position

PI position at UVic

Faculty

2. Principal applicant (students & post-docs)

For further information about the distinction between the Principal Investigator and Principal Applicant, please see the annotated guidelines.
A Principal Applicant is an undergraduate student, graduate student or post-doctoral fellow who will be the lead researcher (for their thesis, dissertation, project, etc.) for this study. A Principal Applicant will be granted "View and edit" access by default, and will receive notifications related to the study. If the project has more than one Principal Applicant, the additional individuals should be listed under section A.3 Research Team Members.

Does this application have a principal applicant (UVic student or post-doc conducting this research for their academic degree)?

Yes

PA name

Cassandra Cupryk

PA email

cassandra.cupryk@gmail.com

PA department

Computer Science COSI

PA position

Master's student

PA phone

4632017203

PA graduate secretary's email (if the principal applicant is a graduate student. Leave blank otherwise.)

gradsec@csc.uvic.ca

Is the principal applicant conducting this research for their academic degree at UVic?

Yes

3. Research team members

Individuals and organizations involved in conducting your research. This includes co-principal investigators, additional principal applicants, co-investigators, other UVic students, assistants (paid or unpaid), community organizations, and clients. Team members listed will have "no access" to application as a default. You cannot assign access to team members without Netlink ID. If they need a Netlink ID go to the Affiliate Identity Management System and click on the 'Sponsor' tab to start the process. Once you get the Netlink ID you have to re-enter their name and give access permission to the application.

List all current research team members (including any UVic students or research assistants who will use the received data or biological materials to fulfill UVic thesis, dissertation, or academic requirements) and assign level of access to the application. Inclusion here satisfies only UVic institutional requirements. If you grant "View and Edit" access to more than one person, be aware that the system will not notify users if and when others are making edits to the application.

DO NOT add the PI or PA to this table as that will cause technical permission issues.

Access: ✎ View and edit project 👁 View only ✉ Receive notifications $ Contribute funding

Name	Email	Role in the project	Institutional affiliation	✎	👁	✉	$
Enrique Larios	elarios.vargas@gmail.com	Professor Collaborator, help with research (planning the methodology, theme coding, etc.)	Uvic	☐	☐	☐	☐
Cassandra Petrachenko	cpetrach@uvic.ca	Staff Collaborator: lab manager, helping with data analysis	Uvic	☑	☐	☐	☐
Alessandra Milani	amilani@uvic.ca	Student Collaborator: helping with research (theme coding, etc.)	UVic	☐	☐	☐	☐
Paul Ralph	paulralph@dal.ca	Collaborator	Dalhousie University	☐	☐	☐	☐

B. Project information

1. Project title

Title for your research project. You may not submit two applications with the same title.

Evaluating A Software Engineering Empirical Standards Tool

2. Anticipated duration of the project

a. Anticipated start date for recruitment/data collection

The approximate start date to begin recruitment and data collection for your project should take into account the time it will take to complete and submit this application form and the period of four to six weeks required for ethical review. It is a violation of University of Victoria policy to begin recruitment and data collection before receiving HREB ethics approval.

Upon Approval

b. Anticipated end date for your research project

An approximate end date for recruitment and data collection.

December 2022

3. Is this application linked to one that has been recently submitted to the UVic Human Research Ethics Board?

No

4. Geographic location(s) of the study

Victoria, British Columbia

5. Keywords to categorize your research

Software Engineering | Empirical Research | ☐ | ☐

C. Project funding

1. Have you and/or research team members (their names must be listed under section A. Research team) applied for or been awarded funding for this project?

This information is used to permit the release of funds and to ensure proper reporting of research ethics approval to funding agencies. Please ensure the information in this table is correct.

No

2. Will this project receive funding from the US National Institute of Health (NIH)?

No

3. If you are a faculty member and have indicated above that you have applied for external funding, have you submitted a Research Application Summary Form to the Grants or Contracts unit in the Office of Research Services?
You must submit a research application summary form to the grants or contracts office every time you apply for external funding. Provide explanation, if you haven't done so.

Not applicable

Comments

D. Multi-jurisdictional research

1. Will this research be conducted under the auspices of another academic institution or health authority in BC (e.g. recruiting through their sites/departments/listservs/poster placement, etc.; involving staff, patients, health records; research team member affiliation)? The checklist may be useful, if unsure.
Research Ethics BC harmonized review (a single coordinated review with the other institution(s) listed) applies if A) you will be conducting research under the auspices of any of the institutions listed in REBC website (involving staff, patients, health records, sites and/or recruitment through their sites, including recruitment via poster placement), and/or B) when members of your research team consist of faculty, staff and students from any of the REBC institution(s).

No

2. Does the proposed research require Research Ethics Board (REB) approval from outside BC?

No

3. If this is a multi-jurisdictional research, please indicate your role in the research project (Check all that apply).
If you answered "Yes" to question D.1 please STOP completing this form and contact HRE office ethics@uvic.ca, 250-472-4321 or 250-472-4545 as soon as possible.

☐ Recruiting Participants

☐ Collecting data

☐ Analyzing data (with or without identifiers collected by you and/or your UVic research team members)

☐ Analyzing data that contain identifiers: data to be collected by non-UVic research team members as outlined in this application

☐ Analyzing data that does not contain identifiers: data to be collected by non-UVic research team members as outlined in this application

☐ Dissemination of results via publications, reports, conferences, internet, etc.

☐ Other

4. Additional information

There is no multi-jurisdictional research.

Recruitment of students will be done on Twitter.

E. Other approvals and consultations

1. If additional request(s) for permission/approval are required please complete the section below (check all that apply)

Other approvals and consultations	Yes, approval uploaded	Yes, will provide as received	No approval required
a. School district, superintendent, principal, teacher			☑
b. Health authorities outside BC involving staff, patients, health records, sites and/or recruitment through their sites (including recruitment via poster placement)			☑
c. Other regional government authority			☑
d. Community group (e.g. formal organization, informal collective)			☑
e. UVic Biosafety Committee approval			☑
f. Other approval			☑

Please upload proof of having made request(s) for permission or any permission/approval documents that you received. Please forward approvals upon receiving them. Be assured that ethics approval may be granted prior to receipt of external approvals.

Comments

F. Scholarly review

1. What type of scholarly review has this research project undergone?

☐ External peer review (e.g. granting agency)

☑ Supervisory committee or supervisor - required for all student research projects

☐ None

☐ Other

G. Researcher(s) qualifications

1. In light of your research methods, the nature of the research, and the characteristics of the participants, what training, qualifications, or personal experiences do the principal investigator, the principal applicant, and/or your research team members have?
E.g. research methods course, language proficiency, committee experience, training on the equipment to be used.

Cassandra Cupryk, the principal applicant, has completed a variety of courses focusing on human computer interaction, research strategies, and study design. Her course work has prepared her to carry out the proposed study. This research will be conducted as a part of her Master's Project.

The project supervisor Dr. Margaret-Anne Storey and her collaborator Dr. Paul Ralph are both renowned for their outstanding work in computer science and software engineering research. They both have extensive experience supervising empirical studies.

2. Tri-Council Policy Statement - TCPS2 CORE Tutorial requirements
All UVic graduate students conducting research with human participants for their UVic project, thesis or dissertation are required to complete the Course on Research Ethics (CORE Tutorial) and provide evidence of ethics training by uploading a CORE completion certificate under this section.

List all current UVic graduate students (also listed under A.2 and A.3) involved in this research project for their UVic project, thesis or dissertation, and upload their Course on Research Ethics (CORE) tutorial certificate(s), if available. This CORE certification is required as of September 1, 2020 for new applications - see the human research ethics web page for more information.

Name	Email	Role in the project	CORE tutorial completion date
Cassandra Cupryk	ccupryk@uvic.ca	PA	June 5 2021

Supporting documents

tcps2_core_certificate.pdf *(Other approval, Name: TCPS 2: CORE, Version: Version 1)*; N 18, 2021

Comments

H. Research Involving the First Nations, Inuit and Métis Peoples of Canada

The TCPS2 (chapter 9) is designed to serve as a framework for the ethical conduct of research involving Aboriginal (including First Nations, Inuit and Métis) or Indigenous peoples, regardless of where they reside or whether or not their names appear on an official register. Its purpose is to ensure, to the extent possible, that research involving Indigenous peoples is premised on respectful relationships and encourages collaboration and engagement between researchers and participants.

This Policy acknowledges the role of the community in shaping the conduct of research that affects First Nations, Inuit, and Métis peoples. The nature and extent of community engagement should be determined through discussion with, and under the advisement of, the relevant community, taking into account relevant characteristics and protocols and the nature of the research.

The University of Victoria Indigenous Plan recognizes that research with Indigenous communities or involving Indigenous peoples must be conducted in a respectful and culturally appropriate manner, following protocols regarding entering community sites, engaging with communities, Elders and Knowledge Keepers, acknowledging cultural knowledge and cultural property, and disseminating research findings.

1. Conditions of the research

a. Will you be conducting research that is situated on any of the following kinds of lands or waterways: First Nation reserves, Indigenous settlements, Indigenous lands under self-government agreements, territories with Indigenous land claims agreements, or other lands designated by Federal, Provincial, or local governments as Indigenous territory?

No

b. Do any of the criteria for participation include belonging to an Indigenous nation, community, group of communities, or organization, including urban Indigenous populations?

No

c. Does the research seek input from participants regarding Indigenous cultural heritage, cultural practices, artifacts, Indigenous or traditional knowledges, or distinct characteristics of Indigenous experience or reality?

No

d. Will Indigenous identity or membership in an Indigenous community or group (e.g. Métis Nation) be used as a variable for the purposes of analysis?

No

e. Will the results of the research make specific reference to Indigenous communities, homelands and/or waterways, peoples, languages, histories or cultures?

No

2. Indigenous engagement

a. Processes and protocols for engagement differ across communities, organizations, committees, and groups, as well as across different research contexts. Describe the process that you have followed with respect to Indigenous engagement.
Include any documentation of collaboration (e.g. formal research agreement, letter of approval, email communications, advisory committee, mentorship, etc.) and the role or position of those consulted (e.g. Elder, Knowledge Holder, governing body, Chief, etc.), including their names, if appropriate.

Not applicable.

b. Explain how Indigenous community members will be meaningfully involved throughout the research process, from research design to knowledge sharing.
Outline the plan, as developed with the community, for the outcomes of the research, including research data ownership, sharing, storage, and governance.

Not applicable.

c. If you have answered "yes" to any of the questions in H.1 but have not yet engaged with the community, committee, organization, or group, please explain why not and outline how you plan to conduct a study that respects Indigenous communities and participants in the absence of prior engagement.

3. Comments

I. International research

1. Will this study be conducted in a country other than Canada?

No

J. Description of research project

1. Briefly describe in non-technical language

a. The research objective(s) and question(s)

We aim to understand whether the Empirical Standards Tool can help Computer Science Ph.D and Master's students students better review research papers.

b. The importance and contributions of the research

The Empirical Standards tool is a website composed of multiple online checklists with essential information that would provide guidance to Computer Science Ph.D and Master's students when they are reviewing research papers.

The major contribution of this research is to discover whether the Empirical standards tool can help Computer Science Ph.D and Master's students review research papers. This information can then be used to improve the Empirical Standards tool so that the tool can better serve Computer Science Ph.D and Master's students.

c. If applicable, provide background information or details that will enable the Research Ethics Board to understand the context of the study when reviewing the application

This research will be conducted as part of a Master's Project of the principal applicant at the University of Victoria. The research information will be accessed only by the researchers.

K. Recruitment

1. Participant details

Provide details of your participants

a. Briefly describe the target population(s) for recruitment
Ensure that all participant groups are identified (e.g. group 1 - teacher, group 2 - administrators, group 3 - parents).

Computer Science Master's and Ph.D. students

b. Why is each population or group of interest?

The tool is aimed to primarily help Computer Science Master's and Ph.D. students who are currently looking to improve how they review research papers.

c. What are the salient characteristics of the participants for your study (e.g. age, gender, ethnicity, class, position, etc.)?
List all inclusion and exclusion criteria you are using.

Computer Science Master's and Ph.D. students who are currently looking to improve how they review research papers.

d. What is the desired number of participants for each group?

5-12 participants

2. Recruitment and process

Provide details of your recruitment process

a. List all source for information used to contact potential participants
E.g. personal contacts, listserves, publicly available contact information, etc. Clarify which sources will be used for which participant groups.

Computer Science Master's and Ph.D. Students will be recruited via Twitter.

b. List all methods of recruitment
E.g. in-person, by telephone, letter, snowball sampling, word-of-mouth, advertisement, etc. If you will be using "snowball" sampling, clarify how this will proceed (i.e. will participants be asked to pass on your study information to other potential participants?). Clarify which methods will be used for which participant groups.

Study Protocol

Recall that this study is being conducted under the supervision of Dr. Margaret-Anne Storey.

Before the study

Firstly, we will be tweeting a tweet (Appendix_09_Tweet_of_Study) that includes the information of the study on Cassandra Cupryk's twitter account. The tweet (Appendix_09_Tweet_of_Study) includes a link and QR code to a Microsoft Form (Appendix 06 – Microsoft Form). Any students that are interested in the study will fill out the form. We will then send an email (Appendix 10 – Email to the Participants and Consent Form) as well as a calendar invitation for the study's Zoom meeting to the participants who have filled out the Microsoft Form. In the email/consent form, we will inform the participants that they imply their consent to participate in the study by attending the study on the outlined date and time.

During the study

For the entirety of the online study, we will present the presentation (Appendix_03_Presentation) alongside the script (Appendix_02_Script).

At one point of the study, the participants who came to the study will be sent the survey (Appendix_04_Survey) via email. The emails of the participants will be obtained from the spreadsheet of emails (Appendix_07_Example_of_List_of_Emails_Collected_from_Microsoft_Form) generated from the Microsoft Form (Appendix_06_Microsoft_Form).

First part of the study

For section 1 of the survey, the participants will need to read 2 research papers. They will then use the Empirical Standards tool to review the 2 research papers and then they will download the reviews to their devices. These reviews will be in the form of .txt files. They will then copy and paste the text from the text files to the survey. If the participants forget what to do in Section 1, the video of the demo (Appendix_08_Link_to_Demo_Video_in_Google_Drive) will be linked in Section 1 of the survey (Appendix_04_Survey). Sections 2 to 7 of the survey are a number of questions allowing the participants to express their opinions of the Empirical Standards tool.

Second part of the study
For the second part of the study, the students will have an open discussion with everyone in the study about their opinions of the study.

c. If you will be using personal and/or private contact information to contact potential paticipants (as stated above), have the potential participants given permission for this, or will you use a neutral third party to assist you with recruitment?
Note that this is not a concern when public and/or business contact information is used.

We will rely on a tweet on Twitter to interact with the students.

d. Who will recruit/contact participants?
E.g. researcher, assistant, third party, etc. Clarify this for each participant group.

A tweet on Twitter.

e. List and explain any relationship between the members of the research team (including third party recruiters or sponsors/clients of the research) and the participant(s) (e.g. acquaintances, colleagues)

Complete section 3 (Power relationship) if there is potential for a power relationship or a perceived power relationship (e.g. instructor-student, manager-employee, etc.). If you have a close relationship with potential participants (e.g. family member, friend, close colleague, etc.) clarify the safeguards that you will put in place to mitigate any potential pressure to participate.

The investigators have no known relationship with any of the participants, but may be supervisors of the students or teaching them courses.

f. In chronological order (if possible) describe the steps in the recruitment process
Include how you will screen potential participants, where applicable. Consider where in the process permission of other bodies may be required.

Study Protocol

Recall that this study is being conducted under the supervision of Dr. Margaret-Anne Storey.

Before the study

Firstly, we will be tweeting a tweet (Appendix_09_Tweet_of_Study) that includes the information of the study on Cassandra Cupryk's twitter account. The tweet (Appendix_09_Tweet_of_Study) includes a link and QR code to a Microsoft Form (Appendix 06 – Microsoft Form). Any students that are interested in the study will fill out the form. We will then send an email (Appendix 10 – Email to the Participants and Consent Form) as well as a calendar invitation for the study's Zoom meeting to the participants who have filled out the Microsoft Form. In the email/consent form, we will inform the participants that they imply their consent to participate in the study by attending the study on the outlined date and time.

During the study

For the entirety of the online study, we will present the presentation (Appendix_03_Presentation) alongside the script (Appendix_02_Script).

At one point of the study, the participants who came to the study will be sent the survey (Appendix_04_Survey) via email. The emails of the participants will be obtained from the spreadsheet of emails (Appendix_07_Example_of_List_of_Emails_Collected_from_Microsoft_Form) generated from the Microsoft Form (Appendix_06_Microsoft_Form).

First part of the study

For section 1 of the survey, the participants will need to read 2 research papers. They will then use the Empirical Standards tool to review the 2 research papers and then they will download the reviews to their devices. These reviews will be in the form of .txt files. They will then copy and paste the text from the text files to the survey. If the participants forget what to do in Section 1, the video of the demo (Appendix_08_Link_to_Demo_Video_in_Google_Drive) will be linked in Section 1 of the survey (Appendix_04_Survey). Sections 2 to 7 of the survey are a number of questions allowing the participants to express their opinions of the Empirical Standards tool.

Second part of the study
For the second part of the study, the students will have an open discussion with everyone in the study about their opinions of the study.

Please upload all the supporting documents relevant to the recruitment methods identified in this section
Examples of supporting documents: email recruitment script, poster, invitation letter, etc. Where draft versions are uploaded please ensure that final versions are submitted when available. If final versions differ significantly after you have obtained research ethics approval, you will need to submit a Request for Amendment.

Supporting documents

Appendix_06_Microsoft_Form_CC_v4.pdf *(Recruitment document, Name: A06 - Microsoft Form, Version: v4)*; J 6, 2022

Appendix_07_Example_of_List_of_Emails_Collected_from_Microsoft_Form_CC_v4.xlsx
(Recruitment document, Name: A7 - Ex. Information from Microsoft Form, Version: v4); J 6, 2022

Appendix_09_Tweet_of_Study_CC_v4.docx *(Recruitment document, Name: A9 - Tweet of Study, Version: v4)*; J 6, 2022

3. Power relationship (dual-role and power-over)

If you are completing this section, please refer to the guidelines for ethics in dual-role research for teachers and other practitioners and the TCPS2, article 3.1 and article 7.4.

Are you or any of your co-researchers in any way in a power relationship, including dual-roles, that could influence the voluntariness of a participant's consent? Could you or any of your co-researchers potentially be perceived to be in a power relationship by potential participants?

Examples of "power relationships" include teachers-students, therapists-clients, supervisors-employees and possibly researcher-relative or researcher-close-friend where elements of trust or dependency could result in undue influence.

Varies

Describe below

a. The nature of relationship

Possible teacher-student

b. Why it is necessary to conduct research with participants over whom you have power relationship

The student of the professor may see the advertisement for the study on accident, since the advertisement of the study will be publicly available on Twitter.

c. What safeguards (steps) will be taken to ensure voluntariness and minimize undue influence, coercion or potential harm

I state in the consent form (A10 - Email to Participants/Consent Form), the following:

There are no known or anticipated risks to you by participating in this research. Research participation is voluntary and you are under no obligation to participate. In addition, choosing to participate or not to participate will not affect your grades, your standing, or your relationships, etc. (as appropriate). The study will be completed in an online setting where other Computer Science Master's and Ph.D. students and study investigator(s) will be present.

d. How the power or dual-role relationship and associated safeguards will be explained to potential participants

I state in the consent form (A10 - Email to Participants/Consent Form), the following:

There are no known or anticipated risks to you by participating in this research. Research participation is voluntary and you are under no obligation to participate. In addition, choosing to participate or not to participate will not affect your grades, your standing, or your relationships, etc. (as appropriate). The study will be completed in an online setting where other Computer Science Master's and Ph.D. students and study investigator(s) will be present.

L. Data collection methods

1. Data collection methods

Use the following sections in ways best suited to explain your project. If you have more than one participant group, be sure to explain which participant group(s) will be involved in which activity/activities or method(s).
If this research will/may include in-person activities during the global pandemic, you must fulfill the requirements supporting in-person research with human participants. Please complete relevant section of the application and appendices with the information outlined in the current UVic Human Research Ethics COVID-19 Bulletin, under the human research ethics webpage.

a. Which of the following methods will be used to collect data? Check all that apply

If this research will/may include in-person activities during the global pandemic, you must fulfill the requirements supporting in-person research with human participants. Please complete relevant section of the application and appendices with the information outlined in the current UVic Human Research Ethics COVID-19 Bulletin, under the human research ethics webpage.

☑ i) Interviewing participants

- ☐ In person
- ☐ By telephone
- ☑ Conducting group interviews or discussions (including focus group)
- ☑ Using web-based technology

Explain and provide the name of the web-based technology or technologies (e.g., Skype, Bluejeans, etc.). For more information on platforms, programs, security, etc. when conducting research with participants virtually see the UVic FAQ and the U.S. Freedom Act advisory below.

If using a web program (online surveys, video conferencing etc.) with a server located in the United States (e.g. SurveyMonkey), or if there are other reasons that the data will be stored in the US (e.g. use of US-based cloud technology, sharing data with US colleagues, etc.), you must inform participants that their responses may be accessed via the U.S. Freedom Act. Please add the following to the consent form(s): "Please be advised that this research study includes data storage in U.S.A. As such, there is a possibility that information about you that is gathered for this research study may be accessed without your knowledge or consent by the U.S. government, in compliance with the U.S. Freedom Act."

Audio and video data will recorded via Zoom.
A facilitated discussion will be done during the study and may use software such as Miro, etc.
Google drive will be used in order to store some of the materials for the study and will be accessed by the participants:
- The 2 PDFs for the research papers
- The video demo on how to complete Section 1 of the survey

If this research will/may include in-person activities during the global pandemic, you must fulfill the requirements supporting in-person research with human participants. Please complete relevant section of the application and appendices with the information outlined in the current UVic Human Research Ethics COVID-19 Bulletin, under the human research ethics webpage.

☑ ii) Administering a questionnaire or survey

- ☐ In person
- ☐ By telephone
- ☐ Email
- ☐ Mail back
- ☑ Web-based

Explain and provide the name of the web-based technology or technologies (e.g., SurveyMonkey), and see the U.S. Freedom Act advisory below.
If using a web program (online surveys, video conferencing etc.) with a server located in the United States (e.g. SurveyMonkey), or if there are other reasons that the data will be stored in the US (e.g. use of US-based cloud technology, sharing data with US colleagues, etc.),you must inform participants that their responses may be accessed via the U.S. Freedom Act. Please add the following to the consent form(s): "Please be advised that this research study includes data storage in U.S.A. As such, there is a possibility that information about you that is gathered for this research study may be accessed without your knowledge or consent by the U.S. government, in compliance with the U.S. Freedom Act."

Survey will administered using SurveyMonkey.

- ☐ Other

☑ iii) Administering a computerized task (describe in section L.1b and/or upload documents)

☑ iv) Observing participants. In section L.1b describe who and what will be observed. Include where observations will take place. If applicable, upload an observational collection sheet for review.

☑ v) Recording of participants and data

- ☑ Audio
- ☑ Video

How are the video images going to be used?

- ☑ Images used for analysis
- ☐ Images used in disseminating results (include release to use participant images' in consent materials)
- ☐ Photos or slides
- ☑ Note taking
- ☐ Flipcharts
- ☐ Data collection sheets (upload)
- ☐ Other

Refers to information/data that was originally gathered for a purpose other than the proposed research and is now being considered for use in research (e.g. patient or school records, personal writings, lesson plans, etc.).

☐ vi) Using human samples (e.g. saliva, urine, blood, hair)

☐ vii) Using specialized equipment/machines (e.g. ultrasound, EEG, prototypes, etc.) or other (e.g. testing instruments that are not surveys or questionnaires)

☐ viii) Using other testing equipment not captured under other categories

E.g. artifacts, paintings, drawings, photos, slides, art, journals, writings, etc.

☑ ix) Collecting materials supplied by, or produced by, the participants

Please specify

- We will be asking the students to complete a review for each research paper and upload the reviews to the survey. We'll then collect the surveys completed by the students.

Refers to information/data that was originally gathered for a purpose other than the proposed research and is now being considered for use in research (e.g. patient or school records, personal writings, lesson plans, etc.).

☐ x) Analyzing secondary data or secondary use of data

☐ xi) Other

b. Provide a sequential description of the procedures/methods to be used in your research study
Be sure to provide details for all methods checked in section L.1. Clarify which procedures/methods will be used for each participant group. Indicate which methods, if any, will be conducted in a group setting. List all of the research instruments and interview/focus group questions, and append copies (if possible) or detailed descriptions of all instruments. If not yet finalized, provide drafts or sample items/questions..

If using a web program (online surveys, video conferencing etc.) with a server located in the United States (e.g. SurveyMonkey), or if there are other reasons that the data will be stored in the US (e.g. use of US-based cloud technology, sharing data with US colleagues, etc.),you must inform participants that their responses may be accessed via the U.S. Freedom Act. Please add the following to the consent form(s): "Please be advised that this research study includes data storage in U.S.A. As such, there is a possibility that information about you that is gathered for this research study may be accessed without your knowledge or consent by the U.S. government, in compliance with the U.S. Freedom Act."

Study Protocol

Recall that this study is being conducted under the supervision of Dr. Margaret-Anne Storey.

Before the study

Firstly, we will be tweeting a tweet (Appendix_09_Tweet_of_Study) that includes the information of the study on Cassandra Cupryk's twitter account. The tweet (Appendix_09_Tweet_of_Study) includes a link and QR code to a Microsoft Form (Appendix 06 – Microsoft Form). Any students that are interested in the study will fill out the form. We will then send an email (Appendix 10 – Email to the Participants and Consent Form) as well as a calendar invitation for the study's Zoom meeting to the participants who have filled out the Microsoft Form. In the email/consent form, we will inform the participants that they imply their consent to participate in the study by attending the study on the outlined date and time.

During the study

For the entirety of the online study, we will present the presentation (Appendix_03_Presentation) alongside the script (Appendix_02_Script).

At one point of the study, the participants who came to the study will be sent the survey (Appendix_04_Survey) via email. The emails of the participants will be obtained from the spreadsheet of emails (Appendix_07_Example_of_List_of_Emails_Collected_from_Microsoft_Form) generated from the Microsoft Form (Appendix_06_Microsoft_Form).

First part of the study

For section 1 of the survey, the participants will need to read 2 research papers. They will then use the Empirical Standards tool to review the 2 research papers and then they will download the reviews to their devices. These reviews will be in the form of .txt files. They will then copy and paste the text from the text files to the survey. If the participants forget what to do in Section 1, the video of the demo (Appendix_08_Link_to_Demo_Video_in_Google_Drive) will be linked in Section 1 of the survey (Appendix_04_Survey). Sections 2 to 7 of the survey are a number of questions allowing the participants to express their opinions of the Empirical Standards tool.

Second part of the study
For the second part of the study, the students will have an open discussion with everyone in the study about their opinions of the study.

c. Where will participation take place for each data collection method/procedure?
Provide specific location (e.g. UVic classroom, private residence, participant's workplace). Clarify the locations for each participant group and/or each data collection method.

Participant's workplace, home, or a place where they are comfortable attending the online workshop.

d. For each method, and in total, how much time will be required of participants?
Clarify this for each participant group, each data collection method, and any other research related activities.

3 hours in total:
- Participants reviewing 2 research papers using the tool and submitting the reviews of the 2 research papers.
- Participants completing a survey evaluating the tool.
- Participants having a group discussion evaluating the study.

e. Will participation take place during participants' office work/hours or instructional time?

No

2. Data collection materials checklist

Data collection methods checklist

- ☑ Standardized instrument
- ☑ Survey
- ☐ Questionnaire
- ☑ Interview and/or focus group questions
- ☑ Observation protocols
- ☐ Other

Please make sure that you have uploaded all the documents relevant to this section. Add any other documents that you think may be relevant to this section.
Where draft versions are appended please ensure that final versions are submitted when available. If final versions differ significantly after you have obtained research ethics approval, you will need to submit a Request for Modification.

Supporting documents

Appendix_01_Description_of_Tool_and_Instructions_To_Use_Tool_CC_v3.pdf
(Data collection instrument, Name: A1 - Description & Instructions of Tool, Version: V3); J 7, 2022

Appendix_02_Script_CC_v3.pdf *(Data collection instrument, Name: A2 - Script, Version: V3)*; J 7, 2022

Appendix_03_Presentation_CC_v3.pdf *(Data collection instrument, Name: A3 - Presentation, Version: V3)*; J 7, 2022

Appendix_04_Survey_CC_v3.pdf *(Data collection instrument, Name: A4 - Survey, Version: V3)*; J 7, 2022

Appendix_08_Link_to_Demo_Video_in_Google_Drive_CC_v4.docx
(Data collection instrument, Name: A08 - Link to Demo Video in Google Drive, Version: v4); J 6, 2022

Appendix_05_Protocol_CC_v4.docx *(Data collection instrument, Name: A5 - Protocol of Study, Version: v4)*; J 7, 2022

M. Possible benefits, inconveniences, and risks of harm to participants

1. Benefits

Identify any potential or known benefits associated with participation and explain below
Keep in mind that the anticipated benefits should outweigh any potential risks.

- ☐ To the participants

☐ To society

☑ To the state of knowledge

Please explain

Will provide a deep understanding of whether the tool can help Software Engineering Ph.D students with reviewing other individuals' research papers.

2. Inconveniences

Identify and describe any known or potential inconveniences to participants
Consider all potential inconveniences, including total time devoted to the research.

No inconveniences apart from time commitment of maximum 3 hrs.

3. Level of risk

The TCPS 2 article 6.12 definition of "minimal risk research" is as follows: 'Research in which the probability and magnitude of possible harms implied by participation in the research is no greater than those encountered by the participant in those aspects of their everyday life that relate to the research.'

Based on this definition, do you believe your research qualifies as 'minimal risk research'?

Yes

Explain your answer with reference to the risks of the study and the vulnerability of the participants

Participants are being asked to volunteer around 3 hours of their time.

4. Estimate of risks of harm

Consider the inherent foreseeable risks associated with your research protocol and complete the table below by selecting the options that best fit the potential risks listed below. Be sure to take into account the vulnerability of your target population(s) if applicable.

Potential risks of harm	Very unlikely	Possibly	Likely
a. Emotional or psychological discomfort, such as feeling demeaned or embarrassed due to the research	☑		
b. Fatigue or stress	☑		
c. Social risks, such as stigmatization, loss of status, privacy and/or reputation	☑		
d. Physical risk such as falls	☑		
e. Economic risks (e.g. job security, salary loss, etc.)	☑		
f. Risk of incidental findings (see article 3.4 of the TCPS 2 for more information)	☑		
g. Other risks	☑		

If other risks, please specify

5. Possible risks of harm

If you indicated in item 4 (a) to (g) that any risks of harm are possible or likely, please explain below

a. What are the risks?
I.e. elaborate on risks you have identified above.

b. What will you do to try to minimize, mitigate, or prevent the risks?

c. How will you respond if the harm occurs?
I.e. what is your plan?

d. If you have indicated that there is a risk of incidental findings in item 4 (f), please outline your proposed protocol for information and /or action

e. If one of your participant groups could be considered vulnerable, please describe any specific considerations you have built into the protocol to address this

6. Risk to researcher(s)

Does this research study pose any risks to the researchers, assistants and data collectors?

No

7. Deception

Will participants be fully informed of everything that will be required of them prior to the start of the researcher session?

Yes

If not, complete the Request to use Deception form on the ORS website

N. Incentives, reimbursement and compensation

1.Is there any incentive, monetary or otherwise, being offered for participation in the research (e.g. gifts, honorarium, course credits, etc.)?

Yes

Explain the nature of each incentive and why you consider it necessary
Also consider whether the amount or nature of the incentive could be considered a form of undue inducement or affect the voluntariness of consent. Clarify which participant groups will be provided with which incentives.

$60 Amazon Gift Cards may be offered to this participant group in order to attract more participants to the study.

2. Is there any reimbursement or compensation for participating in the research (e.g. for transportation, parking, childcare, etc.)?

No

3. Explain what will happen to the incentives, reimbursement or compensation if participants withdraw during data collection or any time thereafter
E.g. compensation will be pro-rated, full compensation will be given, etc.

If the participant withdraws from the study they will get 100% of the incentive. This will ensure that participants are not forced or will not feel forced or coerced in any way to complete the study.

O. Free and informed consent

Consent encompasses a process that begins with initial contact and continues through to the end of the research process.
Consult article 3.2 of the TCPS 2 and appendix V of the guidelines for further information.

1. Participant's capacity (competence) to provide free and informed consent

Capacity refers to the ability of prospective or actual participants to understand relevant information presented about a research project, and to appreciate the potential consequences of their decision to participate or not participate. See the TCPS 2, chapter 3, section C, for further information.

Identify your potential participants (check all that apply)

a. Competent

☑ i) Competent adults

☐ ii) A protected or vulnerable population (e.g. inmates, patients)

☐ iii) Competent youth aged 13 to 18

☐ iv) Competent children under 13 (who are able to provide fully informed consent)

b. Non-competent

☐ i) Non-competent adults

☐ ii) Non-competent youth

☐ iii) Non-competent children (young children and/or children with limited abilities to provide fully informed consent)

2. Means of obtaining and documenting consent and/or assent:

Check all that apply
When completing this section make sure that you consider all of your participant groups, upload copies of relevant materials and complete section O3.

☐ Signed consent

☐ Verbal consent

☑ Letter of information for implied consent (e.g. anonymous, mail back or web-based survey)

Upload information letter in section O.5 or section S - see template.

☐ Signed or verbal assent for non-competent participants

☐ Other means

☐ Consent will not be obtained

☐ Signed consent from the parents/guardians for youth/child participants

☐ Information letters for the parents/guardians of youth/child participants

3. Informed consent

Describe the exact steps (chronological order) that you will follow in the process of explaining, obtaining, and documenting informed consent
Ensure that consent procedures for all participant groups are identified (e.g. group 1 - teachers, group 2 - parents, group 3 - students). Be sure to indicate when participants will first be provided with the consent materials (e.g. prior to first meeting with the researcher?). If consent will not be obtained, explain why not with reference to the TCPS 2 articles 3.5 and 3.7.

Study Protocol

Recall that this study is being conducted under the supervision of Dr. Margaret-Anne Storey.

Before the study

Firstly, we will be tweeting a tweet (Appendix_09_Tweet_of_Study) that includes the information of the study on Cassandra Cupryk's twitter account. The tweet (Appendix_09_Tweet_of_Study) includes a link and QR code to a Microsoft Form (Appendix 06 – Microsoft Form). Any students that are interested in the study will fill out the form. We will then send an email (Appendix 10 – Email to the Participants and Consent Form) as well as a calendar invitation for the study's Zoom meeting to the participants who have filled out the Microsoft Form. In the email/consent form, we will inform the participants that they imply their consent to participate in the study by attending the study on the outlined date and time.

During the study

For the entirety of the online study, we will present the presentation (Appendix_03_Presentation) alongside the script (Appendix_02_Script).

At one point of the study, the participants who came to the study will be sent the survey (Appendix_04_Survey) via email. The emails of the participants will be obtained from the spreadsheet of emails (Appendix_07_Example_of_List_of_Emails_Collected_from_Microsoft_Form) generated from the Microsoft Form (Appendix_06_Microsoft_Form).

First part of the study

For section 1 of the survey, the participants will need to read 2 research papers. They will then use the Empirical Standards tool to review the 2 research papers and then they will download the reviews to their devices. These reviews will be in the form of .txt files. They will then copy and paste the text from the text files to the survey. If the participants forget what to do in Section 1, the video of the demo (Appendix_08_Link_to_Demo_Video_in_Google_Drive) will be linked in Section 1 of the survey (Appendix_04_Survey). Sections 2 to 7 of the survey are a number of questions allowing the participants to express their opinions of the Empirical Standards tool.

Second part of the study
For the second part of the study, the students will have an open discussion with everyone in the study about their opinions of the study.

4. Ongoing consent

Will your research occur over multiple occasions or an extended period of time (including review of transcripts)?

No

5. Participant's right to withdraw

Article 3.1 of the TCPS 2 states that participants have the right to withdraw at any time and can withdraw their data and human biological materials.

a. Describe what participants will be told about their right to withdraw from the research at any time (i.e., who to contact and how)
If compensation is involved, explain what participants will be told about compensation if they withdraw. If you have different participant groups and/or different data collection methods, clarify the different procedures for withdrawing as necessary.

Participants will be told through the Email to Partipants/Consent Form (APPENDIX 10 - Email to Participants/Consent Form) that they may stop participating in the study at any point without explanation.

b. What will happen to a person's data if they withdraw part way through the study or after the data have been collected/submitted?
If applicable, include information about visual data such as photos or videos. If you have different participant groups and/or different data collection methods, clarify the different procedures for withdrawing as necessary. Ensure this information is included in the consent documents.

- ☐ Participant will be asked if they agree to the use of their data
- ☑ It will not be used in the analysis and will be destroyed
- ☐ It is logistically impossible to remove individual participant data (e.g. anonymously submitted data)
- ☐ When linked to group data (e.g. focus group discussions), it will be used in summarized form with no identifying information

Please make sure that you have uploaded all the documents relevant to this section. Add any other documents that you think may be relevant to this section.
Where draft versions are appended please ensure that final versions are submitted when available. If final versions differ significantly after you have obtained research ethics approval, you will need to submit a Request for Modification.

Supporting documents

Appendix_10_Email_Invitation_To_Participants_and_Consent_Form_CC_v4.docx
(Consent/assent form, Name: A10-Email to Participants/Implied Consen, Version: v4); J 7, 2022

P. Anonymity and confidentiality

1. Anonymity

Anonymity means that no one, including the principal investigator, is able to associate responses or other data with individual participants.

a. Will the participants be anonymous in the data gathering phase of research?

No

b. Will the participants be anonymous in the dissemination of results (be sure to consider use of video, photos)?

Yes

2. Confidentiality

Confidentiality means the protection of the person's identity (anonymity) and the protection, access, control and security of their data and personal information during the recruitment, data collection, reporting of findings, dissemination of data (if relevant) and after the study is completed (e.g. storage). The ethical duty of confidentiality refers to the obligation of an individual or organization to safeguard entrusted information. The ethical duty of confidentiality includes obligations to protect information from unauthorized access, use, disclosure, modification, loss or theft.

a. Are there any limits to protecting the confidentiality of participants?

Yes, there are some limits to the researcher's ability to protect the confidentiality of participants (check all that apply)

E.g. focus groups. The researcher cannot guarantee confidentiality.

☑ Limits due to the nature of group activities

The nature or size of the sample from which participants are drawn makes it possible to identify individual participants (e.g. school principals in a small town, position within an organization).

☐ Limits due to context

The procedures for recruiting or selecting participants may compromise the confidentiality of participants (e.g. participants are identified or referred to the study by a person outside the research team).

☐ Limits due to selection

E.g. legal or professional.

☐ Limits due to legal requirements for reporting

E.g. when there will be data storage in the United States. When using USA based data instruments and data storage systems researchers are responsible for determining if this applies.

☑ Limits due to local legislation such as the U.S. Freedom Act

☐ Other

b. If confidentiality will be protected, describe the procedures to be used to ensure the anonymity of participants and for preserving the confidentiality of their data (e.g. pseudonyms, changing identifying information and features, coding sheet, etc.)
If you will use different procedures for different participant groups and/or different data methods be sure to clarify each procedure.

The workshop is done within a group setting, so the researcher cannot guarantee confidentiality.

However, after the study, the participant's data will be access restricted, password protected, anonymized and will be kept confidential. The researchers are the only ones who will have access to this data. Any identifying information will be removed or generalized so that it is not possible to link the information back to the participant before the public dissemination of results.

c. If there are limits to confidentiality indicated in section P.2.a, explain what the limits are and how you will address them with the participants
If there are different procedures for different participant groups and/or different data collection methods, be sure to clarify each procedure.

The workshop is done within a group setting, so the researcher cannot guarantee confidentiality. The participants will be made aware that this study is a group setting in the signed consent form, thus, they will know that their confidentiality cannot be guaranteed.

In addition, this research study includes data storage in the U.S. As such, there is a possibility that information about the participant that is gathered for this research study may be accessed without the participant's knowledge or consent by the U.S. government in compliance with the U.S. Patriot Act. The participant will be made aware of this information in the signed consent form.

Q. Data management

1. Use(s) of data

a. What use(s) will be made of all types of data collected (field notes, photos, videos, audiotapes, transcripts, etc.)?

The data will be used to identify the problems that arise when students use the Empirical Standards tool. The problems will be noted in order to make improvements to the tool. The data will be anonymized, analyzed, and then used to describe study findings. As part of our commitment to open science, study instruments and anonymized results will be available online for replication purposes.

b. Will your research data be analyzed, now or in future, by yourself for purposes other than this research project?

No

c. Will your research data be analyzed, now or in future, by other persons for purposes other than explained in this application?

No

2. Commercial purposes

Do you anticipate that this research will be used for a commercial purpose?

No

3. Maintenance and disposal of data

Describe your plans for protecting data during the project, and for preserving, archiving, or destroying all the types of data associated with the research (e.g. paper records, audio or visual recordings, electronic recordings, coded data) after the research is completed:

a. Means of storing and securing data

E.g. encryption, password protected computer files, locked cabinet, separation of key codes from raw data etc.

The participant's confidentiality and the confidentiality of their data will be access restricted and password protected. The participant's information will be anonymized and will be kept confidential. The researchers are the only ones who will have access to this data.

All of the data will be stored on a University of Victoria Microsoft Office 365 OneDrive Account. OneDrive is built on the Microsoft Office 365 hyper-scale, enterprise-grade cloud, which delivers advanced security and compliance capabilities. The data is encrypted in transit and at rest. All OneDrive files in the University of Victoria Microsoft 365 account are stored in Canada, however, some automated processing may occur outside of Canada. After conducting the interviews, the data will be stripped of any confidential information and stored in a private Github Repository. All notes will be taken digitally and stored in the Uvic OneDrive account as well.

In addition, the following materials are stored on a Google Drive and will be accessed by the participant during the study:
• PDFs of 2 research papers
• A video on how to complete one of the sections of the survey

b. Location of storing data

Include location of data-storage servers if using web-based technology.

Electronic Information: All OneDrive files in the University of Victoria Microsoft 365 account are stored in Canada, however, some automated processing may occur outside of Canada. Email correspondence will be saved on the University of Victoria email servers. Data stored on Github may have servers located U.S.A. and may be accessed under the US Patriot Act.

c. Duration of data storage

If data will be kept indefinitely, explain why this is necessary and state whether the data will contain identifiers or links to identifiers.

Five years.

d. Methods of destroying or archiving data

If archiving data, please describe measures to secure or protect the data. If the archiving will involve a third party (e.g. library, community agency, Aboriginal band, etc.) please provide details.

Electronic Information: Data and backups will be deleted from the repository.

4. Dissemination

How do you anticipate disseminating the research results? (check all that apply)

- ☑ Thesis/dissertation/class presentation
- ☑ Presentations at scholarly meetings
- ☑ Internet (students: most UVic theses are posted on 'UVicSpace' and can be accessed by the public)
- ☐ Media (e.g. newspaper, radio, TV)
- ☐ Directly to participants and/or groups involved
- ☑ Published article, chapter or book
- ☐ Other

R. Conflict of interest

1. Apart from a declared dual-role relationship (section K.3), are you or any of the research team members in a perceived, actual or potential conflict of interest regarding this research project (e.g. partners in research, private interests in companies or other entities)?

No

S. List of uploaded documents

Review the document requirements list and the uploaded documents to ensure that you have all the applicable documents. Make sure to remove all duplicates. Upload appendices as individual documents, instead of clustering appendices under one attachments. Incomplete applications and applications with incorrectly uploaded appendices will not be reviewed. You will be notified in this case.

App. version	Section	Descriptive name	File name	Type of document
V2.2	G.	TCPS 2: CORE	tcps2_core_certificate.pdf	Other approval
V2.2	L.	A1 - Description & Instructions of Tool	Appendix_01_Description_of_Tool_and_Instructions_To_Use_Tool_CC_v3.pdf	Data collection instrument
V2.2	L.	A2 - Script	Appendix_02_Script_CC_v3.pdf	Data collection instrument
V2.2	L.	A3 - Presentation	Appendix_03_Presentation_CC_v3.pdf	Data collection instrument
V2.2	L.	A4 - Survey	Appendix_04_Survey_CC_v3.pdf	Data collection instrument
V2.2	L.	A08 - Link to Demo Video in Google Drive	Appendix_08_Link_to_Demo_Video_in_Google_Drive_CC_v4.docx	Data collection instrument
V2.2	K.	A06 - Microsoft Form	Appendix_06_Microsoft_Form_CC_v4.pdf	Recruitmer document
V2.2	K.	A7 - Ex. Information from Microsoft Form	Appendix_07_Example_of_List_of_Emails_Collected_from_Microsoft_Form_CC_v4.xlsx	Recruitmer document
V2.2	K.	A9 - Tweet of Study	Appendix_09_Tweet_of_Study_CC_v4.docx	Recruitmer document
V2.2	O.	A10-Email to Participants /Implied Consen	Appendix_10_Email_Invitation_To_Participants_and_Consent_Form_CC_v4.docx	Consent /assent for

V2.2	L.	A5 - Protocol of Study	Appendix_05_Protocol_CC_v4.docx	Data collection instrument

T. Signatory/Departmental sign-off

Select the Chair/Director/Dean or their designate to sign-off on this application for submission. Once signed-off, the application will be submitted to the Human Research Ethics Board for review.

By signing-off the application, the signatory is affirming that adequate research infrastructure is available for the conduct and completion of this research project.

Signatory name

Sudhakar Ganti

A.12 Ethics - TCPS2 Core Certificate for Cassandra Cupryk

PANEL ON RESEARCH ETHICS
Navigating the ethics of human research

TCPS 2: CORE

Certificate of Completion

This document certifies that

Cassandra Cupryk

has completed the Tri-Council Policy Statement:
Ethical Conduct for Research Involving Humans
Course on Research Ethics (TCPS 2: CORE)

Certificate # 0000722678 Date of Issue: **5 June, 2021**

Bibliography

[1] Fred D. Davis. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3):319, September 1989.

[2] Carolyn Emden and Sandra Schubert. Manuscript reviewing: What reviewers have to say. *Contemporary Nurse*, 7(3):117–124, September 1998.

[3] Ana M. Fernández-Sáez, Francisco P. Romero, and Marcela Genero. SLR-TOOL - A Tool for Performing Systematic Literature Reviews:. In *Proceedings of the 5th International Conference on Software and Data Technologies*, pages 157–166, University of Piraeus, Greece, 2010. SciTePress - Science and and Technology Publications.

[4] Egon G. Guba. Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ*, 29(2):75, June 1981.

[5] Steven Hadfield and Dino Schweitzer. Building an undergraduate computer science research experience. In *2009 39th IEEE Frontiers in Education Conference*, pages 1–6, San Antonio, TX, USA, October 2009. IEEE.

[6] Brigette Hales, Marius Terblanche, Robert Fowler, and William Sibbald. Development of medical checklists for improved quality of patient care. *International Journal for Quality in Health Care*, 20(1):22–30, February 2008.

[7] Martin Host and Per Runeson. Checklists for Software Engineering Case Study Research. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 479–481, Madrid, Spain, September 2007. IEEE.

[8] B.A. Kitchenham, S.L. Pfleeger, L.M. Pickard, P.W. Jones, D.C. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IIEEE Trans. Software Eng.*, 28(8):721–734, August 2002.

[9] Barbara Kitchenham, Dag I. K. Sjøberg, O. Pearl Brereton, David Budgen, Tore Dybå, Martin Höst, Dietmar Pfahl, and Per Runeson. Can we evaluate the quality of software engineering experiments? In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM '10*, page 1, Bolzano-Bozen, Italy, 2010. ACM Press.

[10] Irene Korstjens and Albine Moser. Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing. *European Journal of General Practice*, 24(1):120–124, January 2018.

[11] J T Lightfoot. A different method of teaching peer review systems. *Advances in Physiology Education*, 274(6):S57, June 1998.

[12] Christopher Marshall, Pearl Brereton, and Barbara Kitchenham. Tools to support systematic reviews in software engineering: a feature analysis. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, pages 1–10, London, England, United Kingdom, 2014. ACM Press.

[13] Jefferson Seide Molléri, Nauman bin Ali, Kai Petersen, Nasir Mehmood Minhas, and Panagiota Chatzipetrou. Teaching students critical appraisal of scientific literature using checklists. In *Proceedings of the 3rd European Conference of Software Engineering Education*, pages 8–17, Seeon/ Bavaria Germany, June 2018. ACM.

[14] Jefferson Seide Molléri, Kai Petersen, and Emilia Mendes. An empirically evaluated checklist for surveys in software engineering. *Information and Software Technology*, 119:106240, March 2020.

[15] Paul Ralph, Nauman bin Ali, Sebastian Baltes, Domenico Bianculli, Jessica Diaz, Yvonne Dittrich, Neil Ernst, Michael Felderer, Robert Feldt, Antonio Filieri, Breno Bernard Nicolau de França, Carlo Alberto Furia, Greg Gay, Nicolas Gold, Daniel Graziotin, Pinjia He, Rashina Hoda, Natalia Juristo, Barbara Kitchenham, Valentina Lenarduzzi, Jorge Martínez, Jorge Melegati, Daniel Mendez, Tim Menzies, Jefferson Molleri, Dietmar Pfahl, Romain Robbes, Daniel Russo, Nyyti Saarimäki, Federica Sarro, Davide Taibi, Janet Siegmund, Diomidis Spinellis, Miroslaw Staron, Klaas Stol, Margaret-Anne Storey, Davide Taibi, Damian Tamburri, Marco Torchiano, Christoph Treude, Burak Turhan, Xiaofeng Wang, and Sira Vegas. Empirical Standards for Software Engineering Research. 2020. Publisher: arXiv Version Number: 2.

[16] P K Rangachari and S Mierson. A checklist to help students analyze published articles in basic medical sciences. *Advances in Physiology Education*, 268(6):S21, June 1995.

[17] D R Seals and H Tanaka. Manuscript peer review: a helpful checklist for students and novice referees. *Advances in Physiology Education*, 23(1):S52–58, June 2000.

[18] Margaret-Anne Storey, Neil A. Ernst, Courtney Williams, and Eirini Kalliamvakou. The who, what, how of software engineering research: a socio-technical framework. *Empir Software Eng*, 25(5):4097–4129, September 2020.

[19] R. Wieringa. Towards a unified checklist for empirical research in software engineering: first proposal. In *16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012)*, pages 161–165, Ciudad Real, Spain, 2012. IET.

[20] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in Software Engineering*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[21] Yichen Yang, Tongan Cai, Shuyang Huang, and Jiachen Liu. Text & Vision-Fused Framework for Academic Paper Review. page 8, 2019.