

The Pennsylvania State University  
The Graduate School

**MEASUREMENT, ASSESSMENT, AND IMPROVEMENT OF  
STATISTICAL LITERACY IN RELEVANT CONTEXTS**

A Dissertation in  
Statistics  
by  
Sayali Phadke

© 2022 Sayali Phadke

Submitted in Partial Fulfillment  
of the Requirements  
for the Degree of

Doctor of Philosophy

December 2022

The dissertation of Sayali Phadke was reviewed and approved by the following:

Matthew Beckman  
Associate Research Professor of Statistics  
Dissertation Co-Advisor  
Co-Chair of Committee

David Hunter  
Professor of Statistics  
Dissertation Co-Advisor  
Co-Chair of Committee

Kari Lock Morgan  
Assistant Professor of Statistics

Jonna Kulikowich  
Professor of Education

Dennis Pearl  
Research Professor of Statistics

Ephraim Hanks  
Director of Graduate Studies

# Abstract

This doctoral work discusses three projects which jointly consider assessment, improvement, and the underlying measurement of contextualized statistical literacy. The central role of statistical literacy has been discussed extensively in the statistics education literature [1–15], emphasizing its importance as a learning outcome and in promoting a citizenry capable of interacting with the world in an informed and critical manner. However, little is known about the influence on student learning outcomes associated with student perceptions about context choices (e.g., application domain) in classroom examples, assessment tasks, etc. Therefore, research which can inform and improve the practice of statistics education is of paramount importance.

The first project in this work assessed the level of contextualized statistical literacy - statistical literacy vis-a-vis contexts with personal relevance or significance to the students. Specifically, the context of the ongoing COVID-19 pandemic was considered. Towards this goal, an isomorphic assessment of an existing research-based instrument was developed and piloted. Data from the pilot study were analyzed to compare psychometric properties of the original and the modified assessment, as well as to consider test-takers' responses to these assessments in relation to various respondent demographics, survey responses, and item characteristics.

The second project employed statistical methods for causal inference to analyze data from a curricular experiment. This experiment was designed and implemented with the aim of improving the level of contextualized statistical literacy. It was conducted in a coordinated undergraduate introductory statistics course taught at a large research

university on the east coast of the United States. Pre-test and post-test scores were collected using the assessment instruments discussed in the first project.

The third project was an application of the Cognitive Diagnostic Modeling (CDM) framework. In addition to being one of the first applications of CDM to statistics education, statistical problem-solving being an inherently more complex cognitive task [16] makes this work a novel contribution. The project outlined the cognitive skills underlying statistically literate behavior as measured by the assessment instruments in the first project. Specifically, data from the pilot study were analyzed to investigate whether a context familiarity skill plays a role in respondents' ability to answer items pertaining to relevant contexts correctly. A Q-matrix specifying the skills needed to answer each item correctly was developed in order to analyze data using CDM models.

This work contributes to methodological advances which can support future statistics education research, through a substantive topic of statistical literacy. It demonstrates 1) the development of an isomorphic assessment, 2) design and implementation of a randomized curricular experiment, 3) estimation and interpretation of causal effect of a curricular treatment on the intended outcome measured through a research-based assessment, and 4) application of CDM to a problem in statistics education including the formation of a Q-matrix. This work can inform both research and practice in statistics education, thereby benefiting students of statistics.

# Table of Contents

List of Figures	viii
List of Tables	x
Acknowledgments	xii
<b>Chapter 1</b>	
<b>Introduction</b>	<b>1</b>
1.1 Statistics Education Research . . . . .	2
1.1.1 Role of Contexts in Statistics Education . . . . .	4
1.1.1.1 Transfer . . . . .	4
1.1.1.2 Isomorphic Assessment . . . . .	5
1.1.2 Randomized Experiments in Statistics Education Research . . . . .	7
1.1.3 Causal Inference in Statistics Education Research . . . . .	8
1.2 Measurement in Educational Research . . . . .	9
1.2.1 Measurement Frameworks . . . . .	10
1.2.1.1 Classical Test Theory (CTT) . . . . .	10
1.2.1.2 Item Response Theory (IRT) . . . . .	11
1.2.2 Cognitive Diagnostic Modeling (CDM) . . . . .	12
1.2.2.1 Notation . . . . .	13
1.2.2.2 Types of models . . . . .	14
1.2.2.3 Parameter estimation . . . . .	19
1.2.2.4 Model comparison . . . . .	21
1.2.2.5 Extensions . . . . .	22
1.2.2.6 Applications of CDM . . . . .	25
<b>Chapter 2</b>	
<b>Examining the Role of Context in Statistical Literacy Outcomes using an Isomorphic Assessment Instrument</b>	<b>28</b>
2.1 Introduction . . . . .	28
2.1.1 Role of Context and Transfer . . . . .	29
2.1.2 Isomorphic Assessment . . . . .	30
2.2 Methodology . . . . .	31
2.2.1 Assessment Modification . . . . .	32

2.2.2	Study Design . . . . .	38
2.2.3	Statistical Methodology . . . . .	40
2.3	Results . . . . .	41
2.3.1	Comparing Assessment Instruments . . . . .	41
2.3.1.1	Summary of Assessment Performance . . . . .	41
2.3.1.2	Reliability Evidence and Evidence for Validity Argument . . . . .	44
2.3.2	Comparing Student Performance . . . . .	48
2.4	Discussion . . . . .	52
2.4.1	Limitations . . . . .	53
2.4.2	Implication for Future Work . . . . .	54
2.5	Conclusion . . . . .	55

### Chapter 3

	<b>Effects of Teaching Through Relevant Contexts on Statistical Literacy: Evidence from a Curricular Experiment</b>	<b>57</b>
3.1	Introduction . . . . .	57
3.1.1	Role of Contexts in Statistics Education . . . . .	58
3.1.2	Randomized Experiments in Statistics Education Research . . . . .	59
3.1.3	Causal Inference in Statistics Education Research . . . . .	59
3.2	Methodology . . . . .	61
3.2.1	Sample . . . . .	61
3.2.2	Tools . . . . .	61
3.2.3	Experimental design . . . . .	62
3.2.4	Design of intervention . . . . .	62
3.2.4.1	Choice of relevant topics . . . . .	63
3.2.4.2	Example . . . . .	65
3.2.5	Causal inference methodology . . . . .	65
3.3	Results . . . . .	67
3.3.1	Mixed Effects Models . . . . .	68
3.3.1.1	Causal effects without covariates . . . . .	70
3.3.1.2	Causal effects with covariates . . . . .	70
3.4	Discussion . . . . .	75
3.4.1	Limitations . . . . .	75
3.4.2	Implications for research . . . . .	76
3.4.3	Implications for teaching . . . . .	77
3.5	Conclusion . . . . .	77

### Chapter 4

	<b>Application of Cognitive Diagnostic Modeling to Statistical Literacy</b>	<b>79</b>
4.1	Methodology . . . . .	80
4.1.1	Q-matrix . . . . .	80
4.1.2	Analytical approach . . . . .	82
4.1.2.1	Context Skill and Item Response Probabilities . . . . .	83
4.1.2.2	Context Skill and Skill Prevalence . . . . .	84

4.2	Results . . . . .	85
4.2.1	Context Skill and Item Response Probabilities . . . . .	85
4.2.1.1	DINA models . . . . .	85
4.2.1.2	A-CDM models . . . . .	86
4.2.1.3	Model comparisons . . . . .	88
4.2.2	Context Skill and Skill Prevalence . . . . .	89
4.3	Discussion . . . . .	90
4.3.1	Limitations . . . . .	90
4.3.2	Implications for research . . . . .	91
<b>Chapter 5</b>		
	<b>Conclusion</b>	<b>92</b>
5.1	Limitations . . . . .	93
5.2	Implication for Future Work . . . . .	95
5.2.1	Implications for research . . . . .	95
5.2.2	Implications for teaching . . . . .	98
5.3	Conclusion . . . . .	99
<b>Appendix A</b>		
	<b>Assessment Instrument - MBLIS</b>	<b>100</b>
A.1	Blueprint for MBLIS . . . . .	100
A.2	MBLIS instrument . . . . .	106
<b>Appendix B</b>		
	<b>Additional Results for Chapter 2</b>	<b>132</b>
B.1	Respondent Demographics . . . . .	132
B.1.1	Univariate summaries . . . . .	132
B.2	Assessment Response Summaries . . . . .	134
B.3	Reliability and Validity Evidence . . . . .	136
B.4	Regression Results . . . . .	142
B.4.1	Diagnostic plots for the additive model . . . . .	146
B.4.2	Diagnostic plots for the interaction model . . . . .	150
<b>Appendix C</b>		
	<b>Additional Results for Chapter 3</b>	<b>154</b>
C.1	Descriptive Summaries . . . . .	154
<b>Appendix D</b>		
	<b>Q-matrix for Statistical Literacy</b>	<b>164</b>
<b>Bibliography</b>		<b>167</b>

# List of Figures

2.1	Scree plots of eigenvalues from PCA . . . . .	46
2.2	Comparison of total score (out of 37) . . . . .	49
3.1	Boxplot of gain score for both treatments . . . . .	69
3.2	Love plot for modified lab as treatment . . . . .	72
3.3	Love plot for instrument type as treatment . . . . .	73
A.1	Mean Number of Hours . . . . .	117
A.2	(Mean Words Recalled for Nap Group) - (Mean Words Recalled for Caffeine Group) . . . . .	121
A.3	(Mean Score for Co-habiting group) - (Mean Score for Single group) . . . . .	123
B.1	Comparison of total score (out of 37) - separate panels . . . . .	134
B.2	Item Information Curves - BLIS . . . . .	136
B.3	Item Information Curves - M-BLIS . . . . .	137
B.4	Test Information Function and Standard Error - BLIS . . . . .	138
B.5	Test Information Function and Standard Error - M-BLIS . . . . .	139
B.6	Item Characteristic Curves - BLIS . . . . .	141
B.7	Item Characteristic Curves - M-BLIS . . . . .	142



B.8	Histogram of residuals - Full model in Equation 1 . . . . .	146
B.9	Quartile-quartile plot of residuals - Full model in Equation 1 . . . . .	147
B.10	Fitted values versus residuals plot - Full model in Equation 1 . . . . .	148
B.11	Residuals plot - Full model in Equation 1 . . . . .	149
B.12	Histogram of residuals - Full model plus interactions . . . . .	150
B.13	Quartile-quartile plot of residuals - Full model plus interactions . . . . .	151
B.14	Fitted values versus residuals plot - Full model plus interactions . . . . .	152
B.15	Residuals plot - Full model plus interactions . . . . .	153
C.1	Boxplot of gain score by instructor . . . . .	154
C.2	Boxplot of gain score by lab sections . . . . .	155
C.3	Boxplot of gain score by gender . . . . .	156
C.4	Boxplot of gain score by class standing . . . . .	157
C.5	Boxplot of gain score by prior statistics training . . . . .	158
C.6	Boxplot of gain score by whether a respondent is an international student	159
C.7	Boxplot of gain score by self-reported expected course grade . . . . .	160
C.8	Boxplot of gain score by highest education level of a parent/guardian . .	161
C.9	Scatterplot of gain score and pre-test score . . . . .	162
C.10	Scatterplot of post-test and pre-test scores . . . . .	163

# List of Tables

2.1	Real from real data to Real from relevant data . . . . .	33
2.2	Naked to Relevant . . . . .	35
2.3	Real from real data to Real from relevant data . . . . .	36
2.4	Context changed considerably . . . . .	37
2.5	Implicit assumption changed . . . . .	38
2.6	Difference in proportion of respondents correctly answering each item . .	43
2.7	Performance on testlet items . . . . .	44
2.8	Raw and predicted coefficient alpha values . . . . .	45
2.9	Model summaries for IRT models . . . . .	47
2.10	Summary statistics of total scores . . . . .	49
2.11	Survey questions regarding COVID-19 pandemic . . . . .	50
2.12	Self-reported effect of context on ability to respond to the statistical question	51
3.1	Survey questions regarding various contexts . . . . .	64
3.2	Example lab activity . . . . .	65
3.3	Summaries of gain scores . . . . .	68
3.4	Causal effects . . . . .	70

3.5	Causal effects - with covariates . . . . .	74
4.1	DINA estimates and standard errors . . . . .	86
4.2	ACDM context skill main effect estimates and standard errors . . . . .	88
4.3	Comparing various models for MBLIS . . . . .	88
4.4	Comparing skill prevalence across models . . . . .	89
A.1	Original BLIS blueprint [17] . . . . .	103
A.2	Extended blueprint for MBLIS . . . . .	105
B.1	Gender identification: n = 1253 . . . . .	132
B.2	International student: n = 1253 . . . . .	132
B.3	Class standing: n = 1253 . . . . .	133
B.4	Prior statistics training: n = 1253 . . . . .	133
B.5	Expected course grade: n = 1253 . . . . .	133
B.6	Highest education of parent/guardian: n = 1253 . . . . .	134
B.7	Selected-response table . . . . .	136
B.8	Difficulty estimates based on PC model . . . . .	140
B.9	Results from the full regression model . . . . .	143
B.10	Results from the full regression model with interactions . . . . .	145
D.1	Q-matrix for Statistical Literacy using MBLIS/BLIS . . . . .	166

# Acknowledgments

I would like to first thank my advisors - Matthew Beckman, David Hunter, and Kari Lock Morgan - for their guidance and support throughout my life as a PhD student. They have been generous with not only their expertise, time, and advice, but also their kindness and humanity. I will remain grateful for that. I benefitted greatly from guidance provided by Aleksandra Slavković and Bruce Desmarais during my early years as a graduate student. I would like to thank them as well as my committee members Dennis Pearl and Jonna Kulikowich for their advice. The support and funding from Center for Social Data Analytics, and Burt Monroe, greatly helped with my research endeavors.

Discovering my passion for teaching has been a central driving force that helped me remain in graduate school and solidify my commitment to statistics education practice and research. Patricia Buchanan, Daisy Philtron, Cecil Shelton, Jenny Shook, Larkin Hood, and Mary Ann Tobin were instrumental in this journey, and I am grateful to them. I shifted the focus of my research to statistics education late into the program and not too long ago. My success in undertaking the first full dissertation in statistics education within our statistics department was possible thanks to Matt Beckman's teaching and continued mentorship, as well as his decision to take a chance on me. This work benefitted from departmental support from Ephraim Hanks and Murali Haran, and stimulating discussions with Dennis Pearl and Neil Hatfield. Additionally, the tremendously positive energy and sense of community within the at-large statistics education community was a motivating force all along. My interactions with Laura Le and Andrew Zieffler (University of Minnesota), Laura Ziegler (Iowa State University), and Mine Çentinkaya-Rundel (Duke University) helped my understanding along the way. The SEEDS - Statistics Education: Engagement and Development for Students - network has been an important avenue for sharing ideas and finding my own community in this field of research. My conversations and work parties with Vimal Rao and Chelsey Legacy (University of Minnesota), Elijah Meyer (Montana State University), and many other SEEDS-lings were enriching and instructive. Most importantly, though, I will remain eternally grateful to my dear friend,

Vimal Rao. Vimal was the first graduate student working in statistics education who I interacted with. Through numerous conversations, Vimal has helped shape my thinking about not only statistics education but also the nature of scientific inquiry and that of learning.

I was lucky to have friends - Meredith, Justin, Mauricio, and Ann - whose camaraderie brought support and accountability when working on difficult math problems as well as all the ice cream a PhD student can ask for! I remain grateful to them, as well as to Claire, Samidha, Isaac, Cecil, and Chelsey. My friends Minita, Lata, Manjusha, Abinaya, Alia, and Sudheer stood by me and reminded me of the world outside of graduate school all along. I am thankful for their friendship and patience.

A graduate student's success is supported by many others behind the scenes! I would like to thank the watchful eyes of Kathy Smith, helping hands of Laura Burghard, and speedy help from Stephanie Valente in the statistics department. I want to thank my therapist, Dr. Eva Letwin, who made me healthier and without whom I sincerely doubt I would have succeeded in finishing my doctoral program. This is also true for my lifelong mentor and cheerleader, Avinash Paranjape, who shared his pearls of wisdom and unconditional support throughout this journey. My family - cousins, aunts, uncles, and grandparents - stood with me lovingly and patiently through the years, and I am grateful to all of them, particularly Chaitali, Prajakta, and Vratesh. My local family in State College included Latha Bhushan and Bhushan Jayarao, Asavari and Ninad Pendharkar, and many members of the SIMA (Society for Indian Music and Arts) and Nritya families who I won't name for the risk of missing someone I value very much. Thank you for your love and support.

They say that it takes a village to raise a child, and I believe that to be true for a doctoral dissertation as well! 6000 miles away from the childhood home and family, I survived and thrived in the last eight years owing to the tremendous support and nurturing from my found family here in central Pennsylvania! As a causal inference researcher, it would be irresponsible of me to claim the counterfactual - I do not know how I would have gotten through graduate school without these people. However, I can say with absolute certainty that their presence made this journey more tolerable, healthier, and happier for me. I found a home-away-from-home in my teacher, mentor, and a father figure - Arijit Mahalanabis, and a brother-from-another-mother - Kishan Patel. They have been and continue to be my pillars of support and sources of nourishment, and no words can truly capture my gratitude towards them. I feel the same way about Sindhuja! Without the sisterhood we have formed and share, my life and mental health would be

a lot poorer. Her husband and my dear friend Kevin has helped me many times and continues to support me every step of the way. Rashmi's friendship was critical during difficult moments and treasured during happy ones! Our walks and shared love for eating brought me joy that was essential to surviving a global pandemic when living on one's own. I am also grateful for the unwavering support and love from my partner, Ben, who stood by me during the most exhausting yet exhilarating phases of the dissertation work and helped me see the light at the end of the tunnel. Finally, though she will neither read this nor understand what this is, my dog, Patches, has been instrumental to my sanity and well-being as a graduate student in the last two years.

Last but not at all the least, I want to thank my parents for their steadfast support during this journey! They raised me to be confident and independent, qualities that turned out to be crucial to my success. As I dedicate this thesis to my late father, I remain grateful for my parents' support through the time and the distance.

# Dedication

To my late father, *bābā*,

thank you

for instilling the value of higher education in me,

for always believing in me - more than I did and even when I didn't,

for reminding me to keep dreaming big, and

for always being curious.

# Chapter 1 | Introduction

The importance and role of statistical literacy in our society as well as statistics education has been discussed extensively in the statistics education literature [1–15]. Guiding documents which inform researchers and practitioners alike, such as the GAISE College Report [18], Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) Research Report [19], International Handbook of Research in Statistics Education [1], and GAISE PreK-12 Report [20], highlight this importance vis-a-vis cognitive outcomes, curriculum, teaching practices, and assessments. The American Statistical Association discusses ‘(to) build a statistically literate society’ as one of its objectives under the strategic goal of statistics education [21]. In parallel, the PARIS21 partnership [22] among global organizations including United Nations, European Union, Organisation for Economic Co-operation and Development, International Monetary Fund, and the World Bank also considers statistical literacy to be a focus of its work. Even though definitions of statistical literacy vary in some aspects [23], mainstream conceptualizations of statistical literacy agree that it comprises of a skillset which an individual would benefit from applying to contexts outside of a classroom, a skillset which would allow people to engage with contexts relevant to them from a data-driven point of view. Further, the literature also converges on a firm belief that statistical literacy plays a critical role in promoting a citizenry that is more capable of understanding the world around them, making evidence-based decisions, making sense of statistical insights pertaining to topics relevant to their professional and personal life [24]. This is also critical at a societal level as highlighted by the ongoing COVID-19 pandemic [25, 26].

Naturally, given the importance of statistical literacy in education as well as the society [25], research about statistical literacy ranges from conceptualizations and definitions, assessment among various populations, and efforts to improve the level of statistical literacy. Previous research has considered ways of improving statistical literacy [27–32].



Additionally, recommendations for teaching practices which can improve statistical literacy have been discussed in guiding documents such as the GAISE college report [18, 33] and the GAISE PreK-12 Report [20]. The GAISE college report [18] recommended integrating “real data with a context and a purpose” into statistics instruction. This nudged the community to bring more real datasets into the classroom and teach through examples that may be relevant to students. There is some evidence that this helped students develop a sense for how statistics is relevant to their lives and improve their engagement with and interest in statistics [34]. The statistics education community is not unfamiliar with discussions surrounding the role of contexts in statistics and its instruction. [35] highlighted that there is no statistics without context. With this in mind, this dissertation poses the following three questions. First, are students of statistics able to make sense of statistical insights encountered in their day-to-day lives, especially pertaining to relevant topics? Second, does including relevant contexts in curricular materials cause a differential gain in students’ statistical literacy outcomes? And third, can we apply a diagnostic framework to an assessment of statistical literacy to measure the role of context? These three questions have been turned into specific research questions (RQs) which as listed in Section 1.2.2.6.1 as well as in the chapter which addresses one or more of them. This work contributes to the literature by working with research-based assessments as recommended by [36].

These research questions interact with methodological questions pertaining to assessment development, experimental design, and measurement. The remainder of this chapter is divided into two parts. Section 1.1 considers broader questions pertaining to statistics education research which informed our broader inquiry and Section 1.2 delves deeper into the questions of measurement models and a specific framework - Cognitive Diagnostic Modeling - which was investigated in this work.

A portion of the development of the argument in this introduction is repeated in later chapters.

## **1.1 Statistics Education Research**

The field of statistics education research, though relatively young, is expansive and continues to expand further. An early reflection on the growing field of statistics education is found in [37]. This report summarized the findings from the roundtable organized by the International Association for Statistical Education, including discussions regarding the interaction between research and practice in statistics education. [3] advanced this

discussion and focused on both practice and research in statistics education, highlighting the interaction between the two and advocating for more research-driven practices. In the same year, [38] summarized statistics education projects funded by the National Science Foundation in the recent years then, underlining the research areas those projects interacted with and how they satisfied the first GAISE college report [39]. At the turn of that decade, [40] analyzed work published in *Statistics Education Research Journal* (SERJ), one of the primary venues for publication of statistics education research, during the previous decade (specifically, years 2002-2009). In addition to the background of authors, intended audience, and types of works, they summarized research areas covered by these articles. Reasoning about/understanding of important statistical ideas was the most frequent topic of inquiry. Teaching and Learning related questions were second-most frequent. This analysis concluded with key recommendations encouraging interdisciplinary as well as foundational research.

The report [19] summarizing recommendation from the American Statistical Association's research retreat focusing on statistics education research can be considered seminal in driving research in this area. This report discussed seven key areas of research - Cognitive Outcomes, Affective Constructs, Curriculum, Teaching Practice, Teacher Development, Technology, and Assessment. In addition to a helpful literature review, three components were detailed for each of these areas. 1) research priorities, including specific research questions, central to advancing our knowledge, 2) implications and benefits of answering these research questions, and 3) measurement and assessments tools needed for this research. [41] continued this discussion by summarizing important research-driven developments in statistics education. This work spotlighted several institutional efforts and an international initiative by UNESCO (United Nations Educational, Scientific and Cultural Organization). It concluded with a call for conducting research with the potential impact at the forefront. The *International Handbook of Research in Statistics Education* [1] was another publication instrumental in driving the field, and remains so till date. This handbook discussed foundational topics in statistics education. In addition to solidifying the scope of statistics education research, chapters in this book underscored the importance of conducting holistic research.

In addition to the overarching discussions and recommendations in these reports and other publications, [42] summarized research focused on teaching and learning of statistics, outlining its interaction with teaching practices. Recently, [43] outlined key developments and gaps in statistics education research within client disciplines.

In the remainder of this section, we focus on two specific topics within the statistics

education literature which are most pertinent to the present work.

### **1.1.1 Role of Contexts in Statistics Education**

Considerable amount of work has discussed the value of contexts in statistics education and powerful ways of introducing contexts which are familiar to the students into the curriculum [26,44–48]. [45] highlighted the centrality of contexts to statistics education as well as statistical literacy. [46] laid out the design for an entire course that focuses on real data that can develop statistical modelers and thinkers. [47] posited statistical habits of mind important for learners as well as teachers, the first of them being the role of context in every stage of statistical inquiry. [49] conducted a study which found that contexts played an important role in 10<sup>th</sup> graders' development of inferential reasoning. [44] is a library of datasets and examples which facilitates the inclusion of real datasets into a variety of courses. Finally, [50] synthesized the discussions regarding and highlighted the value of guided inquiry exercises - exercises where multiple questions are built atop the same context. Research has also investigated the relationship between including contexts in curricular materials and students' engagement with and interest in statistics - as well as student outcomes.

Concurrently, studies focusing on measuring and improving statistical literacy among students at various levels have also been conducted [27–32, 51–59]. However, even though previous work has measured statistically literate behavior outside of a classroom setting [60,61], there is limited work proposing research-based assessments of statistical literacy [62–64]. The present work contributes to the literature on assessment of statistical literacy, specifically in relevant contexts.

#### **1.1.1.1 Transfer**

Assuming that applying statistical literacy skills to new contexts would involve a knowledge transfer [65–67], we distinguish statistical literacy skills from the the ability to apply those skills to topics relevant in our lives, and define contextualized statistical literacy. Such a transfer, though it is central to the purpose of statistical literacy, is not encoded in the definition of statistical literacy. Contextualized statistical literacy is statistical literacy as it pertains to relevant contexts, where relevant contexts are conceptualized as ones that are societally relevant at a given time and people would have engaged with and thought about on their own. The key contribution of this work is in creating an instrument to measure contextualized statistical literacy using an existing research-based

assessment of statistical literacy allowing us to examine respondents' statistical literacy skills when they are required to apply those in relevant contexts.

As underscored by [68], the terms near and far transfer are relative and the distance of transfer implied in those terms is open to interpretation. [69] discusses distance vis-a-vis the similarity to problems encountered during instruction. [70] highlight that distance of transfer is an intuitive notion and discuss it as a matter of similarity and familiarity. Near transfer is across contexts students can be expected to be familiar with because they have encountered similar contexts before during instruction or practice. Whereas far transfer involves transfer across contexts which may not be similar, on the surface, to anything students have encountered before. According to [67], any application of statistical literacy skills can be considered to be a transfer problem. However, when considering contextualized statistical literacy, the question of distance of transfer is not straightforward. On one hand, encountering statistical constructs in new contexts increases the distance of transfer. On the other hand, though, irrespective of whether or not these relevant contexts have been introduced in the classroom before, since we conceptualize relevance as familiarity and engagement outside of the classroom, it can be considered to be nearer transfer for a respondent of an assessment of contextualized statistical literacy. When considering this transfer, we must also be also mindful of possible suspension of sense-making [71, 72] whereby familiarity with the context maybe foregone in favor of focusing on the underlying statistical idea. In statistical problem solving, suspension of sense-making can lead to two possible issues. First, it can limit the benefit of introducing familiar context. Second, in responding to items which require an interpretation in-context to choose the correct answer, such suspension can further increase cognitive load.

#### **1.1.1.2 Isomorphic Assessment**

To measure the transfer of statistical literacy skills to relevant contexts, we created an isomorphic version of an existing research-based assessment of statistical literacy. An isomorphic question or item is identical to a base item in structure (concept, phrasing, as well as distractors) and differs only in the context, continuing to measure the same underlying construct [73–75]. Isomorphic items can also be visualized as items with a common base template differing only in context [68]. [76] and [77] refer to these as structural isomorphs to highlight that this framework itself does not guarantee that respondents' cognitive processes in answering these tasks will be comparable. Isomorphic tasks have been studied extensively in the physics education literature [78–83]. Some

work has also been conducted in the computer science education domain [74, 84]. It is worth noting that there is limited work in the statistics education research literature which studies isomorphs. Most of the aforementioned studies deployed isomorphs to gauge learning and understand common misconceptions, and designed the study in such a way that each respondent solved all of the two or more isomorphic problems at different time points in a random order. [78] study was the only exception where each respondent was assigned to one of the two versions of the assessment. Previous work using isomorphic tasks finds that transfer across such tasks is difficult in most circumstances even if only incidental features are switched. There is some evidence, e.g. [79], that more practice on the base topics improves performance as discussed by [67]. The findings in [85] are valuable given the objective of this work. Their work studied transfer across contexts which cross the disciplinary boundaries in which the construct is situated. They studied the effects of algebraic training on contexts within mathematics as well as in physics to find that training in mathematics facilitated transfer to physics but not the other way around. Even though every context in a statistics problem is external to the discipline itself, this work is important to consider because it provides some evidence of facilitating transfer where contextual information has been provided. This would indicate that familiarity with relevant contexts should improve student performance on statistical literacy tasks as compared to an isomorph based on potentially unfamiliar tasks barring any suspension of sense making [71, 72].

These studies of transfer using isomorphic tasks deploy a variety of types of assessments. However, very few of them use research-based assessments. [84] discuss the importance of developing assessment instruments which undergo rigorous process of collecting reliability and validity evidence, and for researchers to adopt these for further research. [86], in outlining ‘a practical approach to validation’ of research-based assessments support the value and importance of this in step 4 - ‘Identify candidate instruments and/or create/adapt a new instrument’ - with a reminder to first look for previously developed instruments. We chose the Basic Literacy in Statistics (BLIS) assessment [17, 64] because of it’s sole focus on statistical literacy and the extensive research conducted to gather reliability evidence and develop a validity argument for its intended use. We created a modification of the BLIS assessment using isomorphic tasks. We refer to the new version as ‘M-BLIS’ hereafter, especially when highlighting comparisons with the original BLIS assessment. Chapter 5.1 describes the development of M-BLIS, and the design and analysis of a study comparing the original and the isomorphic assessment instruments.

## 1.1.2 Randomized Experiments in Statistics Education Research

As recommended by [87], randomized assignment has been discussed in statistics classrooms, starting with introductory curriculum, to highlight its importance in researchers' ability to draw causal inferences. The role of randomized experiments in educational research [88–90] and educational policy evaluations [91–93] has been discussed in the literature. However, controlled experiments can be difficult to conduct in educational settings [1, 40, chap. 3]. Randomization at the student-level can lead to interference [94] requiring randomization at the classroom or school level. Such studies can be resource-intensive due to the requirement of a large number of sections or classrooms as well as teacher training. Very few designed randomized experiments are conducted in statistics education research with a few notable exceptions. [95] randomized students to control and experimental sections to investigate the effect of simulation-based inference curricula. [96] randomly assigned separate mini readings at the student-level to examine the effects of teaching using tools that may be considered to be fun on student learning. [97] implemented a quasi-experimental design wherein two semesters of a given course enrollment were assigned treatment or control to measure the effects of teaching statistics with a critical pedagogy. [98] studied the effect of teaching through Shiny apps by assigning one of the two enrolled course sections into the treatment group.

One of the goals of the present work was to utilize research-based assessments to measure the outcomes of interest, as encouraged in [19]. Three of the four experiments discussed above implemented a similar strategy. [95] used the ARTIST (Assessment Resource Tools for Improving Statistical Thinking) topic scales [99] for specific topics of interest. [96] used two scales measuring attitudes towards statistics - SATS-36 (Survey of Attitudes Toward Statistics, [100]) and SAM (Statistics Anxiety Measure, [101]) and considered pre-test and post-test scores. [97] also gathered pre-test and post-test data on the CAOS (Comprehensive Assessment of Outcomes in Statistics, [102]) and CLES (Constructivist Learning Environment Survey, [103]) scales. Whereas, [98] used course assignments created by the instructional team.

In this work, we conducted a randomized curricular experiment framed to investigate the causal relationship between teaching through relevant contexts and statistical literacy measured using a research-based assessment in an effort to make an evidence-based contribution to the statistics education research literature. It is important to note that a randomized experiment, in the context of this work, differs from a teaching experiment ([104]; used in [105] and [106]) or a design experiment [107]. Chapter 3 discusses the design of the curricular experiment and analyses of data collected from it.

### 1.1.3 Causal Inference in Statistics Education Research

The research questions in Chapter 3 are framed as causal questions, which is a contribution of the present work. Causal conclusions can play a critical role in informing educational practices through a rigorous investigation of ideas that may or may not affect learning [89]. However, examples of research drawing causal conclusion based on well-designed studies are scarce in statistics education literature. Of the experiments discussed above, [95] was the only study which established a causal effect of the treatment (curriculum type) on the learning outcomes. Their work analyzed the data using a multivariate analysis of covariance (MANCOVA) model. [108] used observational data to investigate the relationship between constructivist strategies in the classroom and students' attitudes towards statistics. They discuss using a causal comparative design, however, they warn against drawing causal conclusions due to the analytical strategies used. [109] also conducted an observational study to assess the effect of instructors and instructional practices on student attitudes. The randomized experimental design employed in this work allows for causal interpretation. We analyze data using a multilevel modeling strategy with covariate adjustment. Outside of statistics education literature, some methodological discussions have highlighted the importance and usage of causal inference in educational studies. [110] discussed a method for estimating the causal effect of time-varying instructional treatments. [111] and [92] discussed the importance and implementation of causal-inference-based conclusions in the context of large-scale assessments in education. [112] conducted an extensive survey of various causal inference methodologies and highlight education as an important application area. [113] discussed the role of and ways to improve causal inference in educational research. The present work provides one possible framework for conducting causal analyses for statistics education research, providing a prototype for similar work in the future.

In this work, causal effects are estimated under the potential outcomes framework [114,115]. In this framework, the overall goal is to model the causal effect of a treatment (denoted with  $\mathbf{W}$ ) on an intended outcome or response variable (denoted with  $\mathbf{Y}$ ). In this study, the treatment  $\mathbf{W}$  is the random assignment of a modified curricular component that incorporates relevant contexts (at section level) and outcome  $\mathbf{Y}$  is the gain score which is the change in statistical literacy score from pre-test to post test (at student level). For a particular unit, the causal effect of treatment is the difference in outcomes that would have been observed if a unit was in the treatment group,  $Y_i(1)$ , and the outcome if it was in the control group,  $Y_i(0)$ .  $Y_i(1)$  captures the gain score for a student enrolled in a treated section - a section receiving the modified curricular component, and  $Y_i(0)$

captures the gain score for a student enrolled in a control section - a section receiving the original curricular component intended for the course. These outcomes are referred to as the Potential Outcomes [114, 115]. The unit-level causal effect is  $\tau_i = Y_i(1) - Y_i(0)$  - a difference between the gain score that would be observed if student  $i$  was enrolled in a treated section and the gain score that would be observed if the same student was enrolled in a control section. Each unit has a potential outcome under each of the treatment statuses, and in theory, it is observable. However, we only observe one of them. Classical methods of causal inference assume what is termed the Stable Unit Treatment Value Assumption (SUTVA) by [116]. SUTVA holds that the outcome of the  $i^{th}$  unit depends only on its own treatment status;  $Y_i(\mathbf{W}) = Y_i(W_i)$ . Further, SUTVA states that there is only one version of the treatment. Interest often lies in estimating the Average Treatment Effect (ATE), which is defined as  $\tau = n^{-1} \sum_i \tau_i$ . The estimation of  $\tau$  is not straightforward for designs where the treatment is not randomized at the unit-level. However, as discussed in Section 1.1.2, this is typical for educational studies.

## 1.2 Measurement in Educational Research

Educational assessments are designed to evaluate test-takers' abilities or skills vis-a-vis specific educational outcomes of interest. Such abilities are latent, and well-designed assessment instruments allow for the measurement of such abilities. The process of measuring these latent constructs involves measurement models. These models, despite their statistical underpinnings, are different from other statistical models in that the estimation of the parameters of such models is an essential part of fine-tuning the measurement itself. For example, the assessment of statistical literacy discussed in Chapter 2 assumes that the items on those instruments accurately capture the manifestations of the latent skill - statistical literacy. Any measurement models used to analyze response data from these assessments can and should be used to calibrate the measurement instrument itself. Therefore, a measurement model takes the input of response data - responses of test-takers on items included in the assessment instrument being deployed. One of the outputs of such models, among others, is a simultaneous estimate of the level of latent ability/skill for each test taker and the level needed to answer a given item correctly. The latency of underlying constructs being measured by these instruments and models is an inherent challenge as well as an opportunity for estimation procedures such as the EM algorithm to be applied to these estimation problems. We begin by reviewing two important measurement models in educational



psychology before introducing the framework deployed in the present research. It is important to remember that measurement models in educational psychology are primarily developed to inform test development and calibration. Therefore, their usage purely for retrofitting to existing assessment data should be treated with care.

### 1.2.1 Measurement Frameworks

Much of the introduction to Classical Test Theory (CTT) and Item Response Theory (IRT) frameworks presented here is based on [117]. When discussing the following measurement models, we assume that we are working with an assessment that only includes multiple-choice type items which are scored dichotomously (correct or incorrect).

#### 1.2.1.1 Classical Test Theory (CTT)

Under the Classical Test Theory (CTT), total observed score on an assessment is modeled, relating an individual's total score (response variable) to the same individual's value/location on the latent continuum for the construct of interest. The observed score is typically the raw total number of correct responses. The simplest version of a CTT model is the true score model specified as

$$X_i = \tau_i + \epsilon_i, \tag{1.1}$$

where  $X_i$  is the observed total score for individual  $i$ ,  $\tau_i = \mathbb{E}(X_i)$  - the expectation over repeated administrations of the same instrument to individual  $i$ , and  $\epsilon_i$  is the error term. However, the “true” score is interpreted as the trait score or the measurement of the latent trait of interest for the individual. Three key assumptions of CTT focus on the correlations between trait scores errors. They specify that - 1) errors from a given assessment instrument are uncorrelated with the underlying trait scores on the same instrument, 2) errors from a given assessment instrument are uncorrelated with the underlying trait scores on on a different instrument, and 3) errors from a given assessment instrument are uncorrelated with errors from a different instrument. It is important to note that gain scores, which may be used as an outcome variable to investigate the effects of a new curricular or pedagogical strategy, violate these assumptions since errors on the pre-test and the post-test may be correlated.

In our application, CTT would try to estimate a respondent's level of statistical literacy based on their total score on BLIS or MBLIS, depending on the type of statistical literacy of interest.

### 1.2.1.2 Item Response Theory (IRT)

Item Response Theory (IRT) is a measurement model that uses responses to items (or questions) on assessments to perform a measurement of a latent and continuous variable i.e. respondent's ability in terms of whatever the assessment claimed to measure, for example, statistical literacy. This framework was a paradigm shift from CTT in that instead of modeling the total score as a manifestation of a true score, responses on individual items were modeled. The similarity, however, is in the treatment of the latent ability or construct as a continuous measurement. In an IRT model, the person ability parameters are measured along the same continuum as the item difficulty parameters. The simplest IRT model - the 1PL (one parameter logistic) model - models the probability of a given individual  $i$  answering a dichotomized item  $j$  correctly solely based on the distance between person  $i$ 's ability level ( $\theta^i$ ) and item  $j$ 's difficulty level ( $\delta_j$ ), both measured along the same dimension of interest. Therefore,

$$P(x_{ij} = 1|\theta^i, \delta_j) = \frac{e^{\alpha(\theta^i - \delta_j)}}{1 + e^{\alpha(\theta^i - \delta_j)}}, \quad (1.2)$$

where  $\alpha$  is a discrimination parameter that captures how well an item discriminates between respondents with different ability levels. In a 1PL model,  $\alpha$  is a constant, assuming that all items on an assessment instrument discriminate equally well among respondents at all ability levels. When  $\alpha = 1$ , the 1PL model is then the Rasch model. A 2 parameter (2PL) model estimates item discrimination parameters in addition to difficulty/ability. Finally, a 3PL model includes a parameter for guessing.

Three key assumptions are underlying an IRT model - 1) unidimensionality of the underlying construct being measured, 2) conditional independence of a person's responses to individual items given the ability level, and 3) that the data follow the functional form specified by the model. Multidimensional IRT (MIRT) modeling framework has been developed and discussed in the literature [118] for problems where the model cannot be assumed to provide a sufficiently accurate representation of the two or more dimensions which may be contributing to the responses.

In case of our application, the 1PL model will estimate individual's level of statistical literacy (ability) as well as the level of statistical literacy needed to answer each item correctly (difficulty) by modeling responses to individual items separately.

## 1.2.2 Cognitive Diagnostic Modeling (CDM)

Educational assessment data are often analyzed under the CTT or the IRT framework. These frameworks provide information about test takers' ability locations along a continuum. These analyses can be most helpful for test development or in instances where the underlying construct of interest has been condensed to one continuum. However, if the investigator's purpose is to provide tools for finer-grained feedback to students or tailor curricular and assessment decisions based on estimated student abilities, CTT or IRT estimates may not provide sufficient information about specific cognitive skills. CDM is a measurement framework which relies on expert-specified multidimensional discrete skills to diagnose mastery or the lack of it based on assessment responses. This framework can offer refined information on individual skill profiles in terms of mastery or non-mastery of important skills using statistical analysis of assessment responses. This methodology has some roots in Latent Class Analysis.

A cognitive diagnostic model assesses test takers' ability vis-a-vis latent cognitive skills (LCSs) which are required to answer test questions. The inputs into this model are 1), a dichotomous item-response matrix based on observed responses ( $X_{ij}; i = 1, 2, \dots, I$  persons, and  $j = 1, 2, \dots, J$  items) and a binary Q-matrix [119] ( $Q_{jk}; j = 1, 2, \dots, J; k = 1, 2, \dots, K$ ) specifying whether skill  $k$  is needed to answer item  $j$ . The model has two key components. The first component is the Item Response Function (IRF) specifying the probability for person  $i$  to answer item  $j$  correctly depending on the skills needs for the item and possessed by the person. The second component is the Joint Attribute Distribution (JAD) specifying the joint distribution of all skills specified in the Q-matrix. Through a variety of parameters estimated from this model, the key focus is three quantities - 1) skill distribution i.e. the proportion of respondents with a given skill, 2) skill class distribution i.e. the proportion of respondents possessing a specific combination of all skills, and 3) individual skill profiles i.e. which skills does a given respondent possess.

The aim of a cognitive diagnostic model is to ensure that the test can provide diagnostic feedback on their strengths and weaknesses on these skills. In general, it can be beneficial to think of LCSs as attributes [120] because in their broad capacity, CDMs are also utilized for psychological health assessments. [121] mention that an "attribute may include procedures, heuristics, strategies, skills, and other knowledge components." In those cases, a diagnosis of whether an individual possesses a certain attribute or not can be useful for diagnostic purposes. However, for the purpose of this work which focuses on an assessment of statistical literacy, we will continue to use the term skill. As mentioned earlier, the inputs into this model are 1), a dichotomous item-response matrix

based on observed responses ( $X_{ij}; i = 1, 2, \dots, I$  persons, and  $j = 1, 2, \dots, J$  items) and a binary Q-matrix [119] ( $Q_{jk}; j = 1, 2, \dots, J; k = 1, 2, \dots, K$ ) specifying whether skill  $k$  is needed to answer item  $j$ .

General framing of a CDM assumes that the Q-matrix specifies mastery or non-mastery of a skill. In part, this modeling framework was conceptualized exactly for this reason - to propose an alternative to Item Response Theory (IRT) type frameworks where the latent skill (ability) is measured along a continuum. We now present details of G-DINA [122] - the Generalized Deterministic Input, Noisy “And” gate model. This is a generalization of the DINA [123] - the Deterministic Input, Noisy “And” gate model. Several other commonly used models can also be derived from G-DINA as special cases and we introduce them as well. One of them, the A-CDM (Additive CDM) will also be used in this paper. However, it is important to note that two additional general frameworks for diagnostic modeling are also available and discussed in the literature. [124] introduced the General Diagnostic Model (GDM) and [125] introduced the Loglinear Cognitive Diagnostic Model (LCDM). Both these models express log-odds of correct responses in terms of a linear function of required skills. The original form of the GDM focused on main effects of each skills, whereas LCDM included skill interactions as well. All these frameworks fall under the general umbrella of Cognitive Diagnostic Models. [125, 126] discuss equivalence relationships between models across the three generalized frameworks.

### 1.2.2.1 Notation

We begin by clarifying the notation used for all CDMs discussed in the present work. These notations are critical because they have been updated for additional clarity and consistency vis-a-vis the original works discussing the models.

- $N$  people/test-takers/respondents indexed by  $i = 1, 2, \dots, N$
- $J$  items on the assessment indexed by  $j = 1, 2, \dots, J$
- $X_{[NXJ]}$  is the dichotomous observed item response matrix with  $X_{ij}$  denoting the correctness of person  $i$ 's response on item  $j$ .  $X_{ij} = 1$  indicates a correct response and  $X_{ij} = 0$  an incorrect one.
- $\eta_{ij}$  denotes the latent item response of person  $i$  on item  $j$ .  $\eta_{ij}$  are unobserved.
- $K$  skills/attributes indexed by  $k = 1, 2, \dots, K$

- $Q_{[JK]}$  is the expert-specified dichotomous matrix indicating whether skill  $k$  is required to answer item  $j$  correctly.
- $\alpha^c_{[K]}$  specifies the  $c^{th}$  configuration of skills or a skill profile and is a vector of length  $K$ . There are a total of  $2^K$  possible skill profiles i.e.  $c = 1, 2, \dots, 2^K$
- $\alpha^i_{[K]}$  is individual  $i$ 's skill profile which is a vector of length  $K$ .
- $\pi_{\alpha^c}$  is the joint attribute distribution i.e. joint probability of any skill profile  $\alpha^c$ .
- $\gamma$  capture the modeling parameters to include  $\delta$ s from the item response function (Equation 1.3) and  $\lambda$ s from the joint attribute distribution (Equation 1.2.2.2.3).  $\gamma'$  refers to parameter estimates from the previous EM iteration.

### 1.2.2.2 Types of models

In this section we describe some of the important CDMs. We begin with a generalized model and then describe other important models while also showing how they are special cases of the general model. We discuss the estimation details only for G-DINA because as discussed in [122] and [127], all other models discussed here can be arrived at through matrix transformation of parameters of the general model.

#### 1.2.2.2.1 G-DINA model

In this section we briefly describe the Generalized Deterministic Input, Noisy ‘‘And’’ gate (G-DINA) model [122]. This model is a generalization of the popularly used DINA model [123] - Deterministic Input, Noisy ‘‘And’’ gate - which includes a ‘deterministic input’ i.e. the ideal item response is deterministic, ‘noisy’ probability of correct response, and an ‘And’ gate specifying that this model is non-compensatory and all skills specified by the Q-matrix are required to answer a given item correctly. G-DINA generalizes this non-compensatory constraint to incorporate the relationship between presence of skills in all possible combinations, and considers general specifications of both the IRF as well as the JAD. This model is also generalized to encompass a variety of link functions in describing the probability for a person  $i$  with skill profile  $\alpha^c$  to succeed on item  $j$  and can be easily manipulated to derive other widely used CDMs such the DINA (Deterministic Input, Noisy ‘‘And’’ gate) model [123, 128], the DINO (Deterministic Input, Noisy ‘‘Or’’ gate) model [129], or the additive CDMs. Details of these generalizations are discussed in [122] and [127]. These papers also discuss the relationships between a large variety of CDMs and G-DINA.

[122] proposed a generalized DINA model generalized to incorporate a general specification for the relationships between mastery or non-mastery of a skill required to answer an item correctly and the responses on the items themselves. The canonical form of G-DINA uses the identity link. In our presentation, we modify and clarify some of the notation without any loss of information. The estimation procedure expands on the details provided in [122] and [127] to layout the derivations of an EM algorithm for parameter estimation. Examinees are assumed to be independent in this model.

### 1.2.2.2.2 Item response function (IRF)

The IRF shown in Equation 1.3 specifies the probability that individual  $i$  with attribute profile  $\alpha^c$  ( $a^i = \alpha^c$ ) answers item  $j$  correctly as a function of main and all interaction effects of possessing any of the skills specified in the Q-matrix. The  $g$  is the link function which is identity in the canonical form. However, the derivations using log or logit function are also straightforward. When proposing G-DINA, [122] use a reduced attribute vector for each item  $j$ ,  $K_j$ . Despite the role of the Q-matrix in conceptualizing the CDM framework, the G-DINA model specification does not explicitly utilize the Q-matrix. The preemptive subsetting of attributes also leads to additional assumptions of skill ordering. However, this can be avoided by multiplying the appropriate row of the Q-matrix. This approach clarifies the notation better, is generalized completely, and does not affect the estimation procedure in any way. Equation 1.3 specifies the IRF.

$$g[P(X_{ij} = 1 | a^i = \alpha^c)] = \delta_{j0} + \sum_{k=1}^K \delta_{jk} (Q_{jk} \times \alpha^c_k) + \sum_{k'=k+1}^K \sum_{k=1}^{K-1} \delta_{jkk'} (Q_{jk} \times \alpha^c_k) (Q_{jk'} \times \alpha^c_{k'}) + \dots + \delta_{j12\dots K} \prod_{k=1}^K (Q_{jk} \times \alpha^c_k), \quad (1.3)$$

where  $\delta_{j0}$  is the intercept term for item  $j$ ,  
 $\delta_{jk}$  is the main effect due to having mastered the  $k^{skill}$  on its own,  
 $\delta_{jkk'}$  is the interaction effect due to having mastered subsets of 2 skills jointly,  
and  $\delta_{j12\dots K^*_j}$  is the interaction effect of all skills required for item  $j$ .

### 1.2.2.2.3 Joint attribute distribution (JAD)

The joint attribute distribution  $\pi_c$  specifies the probability distribution for attribute profile  $\alpha^c$ . This is a function of  $\lambda$ s, which are the structural parameters of the JAD. In its

simplest specification under the assumption of independence of attributes, the parameter space includes separate probabilities of possessing each skill. In the simple case,

$$\pi_c = \prod_{k=1}^K \lambda_k^{\alpha_k^c} [1 - \lambda_k]^{1-\alpha_k^c}, \quad (1.4)$$

where  $\alpha_k^c = 1$  indicates that under skill profile  $\alpha^c$ , the  $k^{th}$  skill is mastered. The parameter  $\lambda$ s in this case are the probabilities of mastering each skill -

$$\lambda = [P(\alpha^c_1 = 1), \dots, P(\alpha^c_k = 1), \dots, P(\alpha^c_K = 1)]. \quad (1.5)$$

Another example of a simple parametrization is the saturated model. The saturated model has  $2^K$  values and  $2^K - 1$  parameters -  $\nu = [\pi_1, \dots, \pi_{2^K-1}]^T$  - one for each possible combination of  $K$  skills minus one, capturing the joint probability of each skill profile. These are also known as the mixing proportion parameters and have a constraint that they must sum up to 1 -  $\sum_{c=1}^{2^K} \pi_c = 1$ . Hence the  $2^K - 1$  parameters instead of  $2^K$ . Additionally, more complex attribute distributions and their implications are also discussed in the literature.

#### 1.2.2.2.4 DINA model

DINA - the Deterministic Input, Noisy "And" gate model - was introduced by [130] and is a widely used CDM. DINA stipulates that individual  $i$  can answer question  $j$  correctly only if the individual possesses all the skills specified as per  $Q_j$  - the  $j^{th}$  row of the Q-matrix. This is a non-compensatory model i.e. a respondent who lacks any of the skills required for a given item cannot compensate for that deficiency in any other way. For respondent  $i$  and item  $j$ , the latent dichotomous response  $\eta_{ij}$  can be written as,

$$\eta_{ij} = \prod_{k=1}^K [a^i_k]^{Q_{jk}}, \quad (1.6)$$

where  $\eta_{ij}$  is the latent/unobserved response,  $\mathbf{a}^i$  is person  $i$ 's skill composition and  $\mathbf{Q}_j$  is item  $j$ 's skill requirement. Both  $a^i_k$  and  $Q_{jk}$  are binary quantities specifying whether person  $i$  possesses skill  $k$  or not, and whether item  $j$  needs skill or not, respectively. Therefore,

$$\begin{aligned} \eta_{ij} &= 1 \text{ if person } i \text{ possesses all skills required for item } j \\ &\quad \text{(can have more but not fewer) ,} \\ \eta_{ij} &= 0 \text{ otherwise.} \end{aligned} \tag{1.7}$$

However,  $\mathbf{a}^i$  are not observed for any individual. Therefore, we have to account for the possibility that  $\eta_{ij}$  does not reflect whether individual  $i$  has truly mastered or not mastered the skills required for item  $j$ . We define slippage as the instance in which the observed response  $X_{ij}$  is incorrect even if the individual has all the required skills, and define guessing as the inverse where  $X_{ij}$  is correct even if the individual does not have the requisite skills. Therefore,

$$\text{Guessing: } g_j = P(X_{ij} = 1 | \eta_{ij} = 0); j = 1, 2, \dots, J, \text{ and} \tag{1.7}$$

$$\text{Slippage: } s_j = P(X_{ij} = 0 | \eta_{ij} = 1); j = 1, 2, \dots, J. \tag{1.8}$$

The guessing parameter for item  $j$ ,  $g_j$ , is the probability that an individual who does not have all the skills required for item  $j$  will still answer it correctly. The slippage parameter for item  $j$ ,  $s_j$ , is the probability that an individual who does possess all the skills required for item  $j$  will give an incorrect answer. DINA is a special case of G-DINA where all but the highest level of interaction is set to zero. In the G-DINA IRF in Equation 1.3, the intercept term,  $\delta_{j0}$ , is the guessing parameter  $g_j$  and the last term,  $\delta_{j12\dots K}$ , which is the coefficient of an interaction between all skills is  $1 - s_j$ , the probability of answering an item correctly if all required skills are possessed. [125] discuss the constraint that  $(1 - s_j) > g_j$  i.e. an examinee who has mastered all required skills has a higher probability of answering item  $j$  correctly than the examinee's probability of answering correctly if all the required skills are not mastered. However, this constraint is not necessarily incorporated into the estimation procedure.

In Equation 1.11,  $\mathbf{a}^i$  was presumed known. However, since skills among individuals are unobserved, we introduce skill profiles or classes  $\alpha$  to specify the combination of skills possessed by an individual. In cases where an individual can possess any combination of the  $K$  skills, there are  $2^K$  possible  $\alpha$ s. This is assumed to be the case for DINA and is referred to as the saturated skill distribution. The saturated model has  $2^K - 1$  parameters in the Joint Attribute Distribution (JAD) -  $\nu = [\pi_1, \dots, \pi_{2^K - 1}]^T$  - one for each possible combination of  $K$  skills minus one, capturing the joint probability of each skill



profile. Additionally, more complex attribute distributions and their implications are also discussed in the literature. Resultantly,

$$P(X_{ij} = 1 | \mathbf{a}^i = \boldsymbol{\alpha}, \mathbf{g}_j, \mathbf{s}_j) = (1 - s_j)^{n_{ij}} g_j^{(1-n_{ij})} \quad (1.9)$$

There are a total of  $2 * J + 2^k - 1$  parameters in a DINA model with one guessing  $\mathbf{g}_j$  and one slippage  $\mathbf{s}_j$  parameter for each of the  $J$  items, and  $2^k - 1$  probabilities of skill profiles,  $\boldsymbol{\alpha}$ s, with the constraint that they must add up to 1. Estimates of these parameters are utilized for further estimating three quantities - 1) skill distribution for each latent skill, 2) distribution of each of the  $2^k$  skill classes or profiles, and 3) individual skill profiles.

We assume local independence ( $\mathbf{X}_{ij} | \mathbf{a}^i \perp\!\!\!\perp \mathbf{X}_{ij'} \forall i$ ) i.e. conditional on an individual's skill profile, their responses on all the items are independent of each other. Additionally, test-takers are assumed to be mutually independent of each other. Provided all this information, we treat  $X_{ij}$ s to be Bernoulli and write the following:

$$P(\mathbf{X}_i | \mathbf{a}^i = \boldsymbol{\alpha}, \mathbf{g}_j, \mathbf{s}_j) = \prod_{j=1}^J P(X_{ij} = 1 | \mathbf{a}^i = \boldsymbol{\alpha}, \mathbf{g}_j, \mathbf{s}_j)^{X_{ij}} [1 - P(X_{ij} = 1 | \mathbf{a}^i = \boldsymbol{\alpha}, \mathbf{g}_j, \mathbf{s}_j)]^{(1-X_{ij})}. \quad (1.10)$$

#### 1.2.2.2.5 DINO model

The Deterministic Input, Noisy "Or" gate model, or DINO, is another popular CDM. This is the compensatory counterpart of the DINA model where the "Or" gate specifies that the probability of answering an item correctly is based on whether at least one of the skills required for that item has been mastered. Mastering each additional skill increases the probability of a correct answer. [129] introduced this model and discussed that this model is most effective in psychological assessments in which diagnosis may depend on the presence or absence of individual attributes. We write the unobserved response  $\omega_{ij}$  as,

$$\omega_{ij} = 1 - \prod_{k=1}^K (1 - a_k^i)^{Q_{jk}}, \quad (1.11)$$

capturing the two groups of respondents - those with at least one of the required skills and those with none of them. DINO also estimates the slippage ( $s_j$ ) and guessing ( $g_j$ ) parameters, similar to DINA, through the item response function,

$$P(X_{ij} = 1 | \mathbf{a}^i = \boldsymbol{\alpha}, \mathbf{g}_j, \mathbf{s}_j) = (1 - s_j)^{\omega_{ij}} g_j^{(1-\omega_{ij})}. \quad (1.12)$$

DINO model can be derived from G-DINA by writing the IRF in Equation 1.3 as  $g[P(X_{ij} = 1 | a^i = \alpha)] = \delta_{j0} + \delta_{jk}(Q_{jk} \times \alpha_k)$  where the intercept term,  $\delta_{j0}$ , is still the guessing parameter  $g_j$ , and the  $\delta_{jk} = -\delta_j k' k'' = \dots = (-1)^{K+1} \delta_{j12\dots K}$  last term,  $\delta_{j12\dots K}$ .

#### 1.2.2.2.6 A-CDM model

The A-CDM or Additive CDM [122] can be thought of as the main effects model in which all interaction terms in G-DINA IRF in Equation 1.3 are set to zero. In this model, mastering each of the required skills for item  $j$  increases the probability of success on that item. The IRF is written as,

$$g[P(X_{ij} = 1 | a^i = \alpha)] = \delta_{j0} + \sum_{k=1}^K \delta_{jk}(Q_{jk} \times \alpha_k) \quad (1.13)$$

When identity link is applied to Equation 1.13, this model is known as the A-CDM. When  $g(\cdot)$  is a logit link, this model is known as the Logistic Linear Model (LLM) [131], and if  $g(\cdot)$  is a log link, the model becomes Reduced Reparametrized Unified Model (R-RUM) [132].

#### 1.2.2.3 Parameter estimation

Estimation of CDM parameters has been discussed at some length in [128] for the DINA model (Section 1.2.2.2.4) and [122] for the G-DINA model (Section 1.2.2.2.1). Maximum likelihood estimates (MLE) of the parameters and covariance of the estimators based on inverse of the Hessian matrix evaluated at the MLE would involve the unobserved individual skill profiles. The expectation-maximization (EM) algorithm [133] would be an effective tool for such an estimation situation. The EM algorithm involves taking an expectation of the complete data loglikelihood ( $l_{comp}$ ) over the missing data in the E-step i.e. the expectation step. The complete data loglikelihood assumes that individual skill profiles,  $a^i$ s, are known. The maximization step, known as the M-step, maximizes results from the E-step over the parameters. The algorithm iterates until convergence. Since  $a^i$ s are assumed to be known, we do not need to sum over all possible skill profiles in  $l_{comp}$ . For the G-DINA model, the complete data loglikelihood is as follows:

$$l_{comp}(\gamma) = \log \prod_{i=1}^N L(\mathbf{X}_i | a^i = \alpha^c) P(\alpha^c) \quad (1.14)$$

$$= \log \prod_{i=1}^N \left( P(\alpha^c) \prod_{j=1}^J L(X_{ij} | a^i = \alpha^c) \right) \quad (1.15)$$

$$= \log \prod_{i=1}^N \left( P(\alpha^c) \prod_{j=1}^J P(X_{ij} = 1 | a^i = \alpha^c)^{X_{ij}} \left( 1 - P(X_{ij} = 1 | a^i = \alpha^c) \right)^{1-X_{ij}} \right) \quad (1.16)$$

$$= \sum_i \log P(\alpha^c) + \sum_{i,j} \left( X_{ij} \log P(X_{ij} = 1 | a^i = \alpha^c) + (1 - X_{ij}) \log \left( 1 - P(X_{ij} = 1 | a^i = \alpha^c) \right) \right) \quad (1.17)$$

[128] claim to implement an EM algorithm to estimate DINA model parameters. However, derivations presented in this work take a direct differentiation of the observed data loglikelihood with respect to the DINA parameters, without the expectation step from the EM. The observed data likelihood ( $l_{obs}$ ), sometimes referred to as the marginal loglikelihood would include a sum over all possible skill profiles  $a^i$  can take because the skill profiles are not observed.

$$l_{obs}(\gamma) = \log \prod_{i=1}^N \sum_{c=1}^{2^K} L(\mathbf{X}_i | a^i = \alpha^c) P(\alpha^c) \quad (1.18)$$

$$= \log \prod_{i=1}^N \prod_{j=1}^J \sum_{c=1}^{2^K} \left( P(\alpha^c) L(X_{ij} | a^i = \alpha^c) \right) \quad (1.19)$$

$$= \log \prod_{i=1}^N \prod_{j=1}^J \sum_{c=1}^{2^K} \left( P(\alpha^c) P(X_{ij} = 1 | a^i = \alpha^c)^{X_{ij}} \left( 1 - P(X_{ij} = 1 | a^i = \alpha^c) \right)^{1-X_{ij}} \right) \quad (1.20)$$

$$= \sum_{i,j} \log \sum_{c=1}^{2^K} \left( P(\alpha^c) P(X_{ij} = 1 | a^i = \alpha^c)^{X_{ij}} \left( 1 - P(X_{ij} = 1 | a^i = \alpha^c) \right)^{1-X_{ij}} \right) \quad (1.21)$$

$$= \sum_{i,j} \log \sum_{c=1}^{2^K} \left( \prod_{k=1}^K \lambda_k^{\alpha^c k} [1 - \lambda_k]^{1-\alpha^c k} P(X_{ij} = 1 | a^i = \alpha^c)^{X_{ij}} \left( 1 - P(X_{ij} = 1 | a^i = \alpha^c) \right)^{1-X_{ij}} \right) \quad (1.22)$$

[122] take a similar approach with G-DINA. However, this work obtains parameter es-

timates by maximizing the marginal loglikelihood directly with respect to the unobserved skill profiles. Unlike an EM, the properties of this algorithm, particularly pertaining to the expected convergence, are not outlined in these works. However, it appears to provide a closed-form solution and would be certainly easier to implement than EM would be. According to [122], standard errors for parameter estimates of G-DINA are calculated using the multivariate delta method.

[127], in laying out the R package *GDINA*, specify that parameters are estimating using an EM algorithm. Further, [127] specify a solution for the E-step which can be matched by taking an expectation of the complete data loglikelihood in Equation 1.17 with respect to the G-DINA parameters. However, the M-step is not outlined in [127] and its solution is difficult to derive in closed form.

#### 1.2.2.4 Model comparison

Studies have investigated best-fitting CDMs by fitting multiple models to single datasets, or to subsets separated by countries, classrooms, or even at item-by-item basis [134–136]. [135] analyzed English as Second Language (ESL) grammar test performance using the Examination for the Certificate of Proficiency in English (ECPE) data by fitting a full LCDM model as a first step. Consequently, various CDMs were fit at the item-level to determine best fit based on Root Mean Squared Error (RMSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) for each item. [134] fit DINA, DINO, NIDO, C-RUM models to TOEFL reading and listening data to determine best model fit using comparable measures as [135]. [136] also fit a variety of CDMs as well as IRT models at the country-level and item-level in analyzing the Trends in International Mathematics and Science Study (TIMSS) 2007 data to find the empirically best-fitting model based on a variety of model comparison metrics.

These models search for different model across the spectrum - parsimonious models (DINA, DINO), main effects models (A-CDM, LLM, R-RUM), and a saturated G-DINA, or even IRT models. Since each of these models may have varying underlying assumptions and differ in substantive interpretations, such model comparisons that rely solely based on empirically finding best fits based on model diagnostics such as AIC, BIC, loglikelihood, or other absolute measures. We need to ask the questions - which assumptions and modeling choices can we make most sense of for the given application? This is the same issue with the ‘finding the best q-matrix’ algorithm.

On the other hand, [137] compared model fits under a variety of Q-matrix and model mis-specifications through extensive simulation studies. [138] considered item-fit under the

G-DINA model and [139] proposed a two-step likelihood ratio test (LRT) for item-level model comparison under G-DINA as well. [140] introduced a framework for dimensionality assessment that considers whether the number of attributes specified in the Q-matrix is appropriate. [141] discussed a Bayesian method for estimating the Q-matrix using priors that capture expert knowledge. Each Q-matrix entry is considered to be a bernoulli and the probability of success is informed by a logistic regression where predictors are expert knowledge about the items. This method can validate which of the underlying attributes may be predicted by experts and which residual remain unexplained. In the same vain as algorithmically discovering the most suitable Q-matrix, [120] proposed a method for validation of an empirical Q-matrix. This work built on the discrimination index for DINA specified in [142]. To calculate this index, two groups of respondents are compared - those with a correct unobserved latent outcome ( $\eta_{ij} = 1$ ) and those with the incorrect outcome ( $\eta_{ij} = 0$ ). These two groups can be thought of as respondents with all the required skills for item  $j$  and respondents without those skills. A correctly specified Q-vector for item  $j$  will maximize difference between success probabilities for the two groups - the discrimination index.

In this work, we focus on developing a fully expert-specified Q-matrix and fitting only those models which are substantively applicable to the problem at-hand. We also argue that model comparisons should consider underlying assumptions and substantive interpretations when choosing the ‘best fitting’ model.

### 1.2.2.5 Extensions

The CDM framework is expansive, and a variety of models and their extensions have been discussed in the literature. We highlight some of the work to capture key features available methodology. Additional developments such as random-effects approach to modeling DINA parameters [143], differential item functioning (DIF) detection using CDM [144], incorporating predictors [145, 146], and working with missing data [147] have also been discussed. Here, we focus on three specific categories of developments.

#### 1.2.2.5.1 Methods Focusing on the Q-matrix

All models discussed here are based on a dichotomous Q-matrix specifying whether mastering the  $k^{th}$  skill is required for answering item  $j$  correctly. The Ordered Category Attribute Coding (OCAC) framework in [148] explored the notion that skills contributing to assessment responses may be specified at or of interest at more than two levels. These multiple levels need not be mastery levels. They may be interpreted as levels of

mastery, various steps along the learning path of a skill, or nested learning outcomes. [148] proposed models which can incorporate expert-specified mixed skill types (binary versus categorical) or an incomplete Q-matrix where levels of a given skill are discovered through response data.

[149] distinguished between expert-defined and data-defined polytomous attributes where a data-defined attribute or skill is discovered through data. However, such discovered skills are not helpful for assessment design since items cannot be designed with the purpose of assessing those levels. Expert-defined polytomous attributes can allow for such test design. This method works with Specific Attribute Level Mastery (SALM) items, building on the OCAC framework proposed by [148]. SALM items are targeted to measure specific levels of attributes. The proposed pG-DINA (polytomous G-DINA) model incorporates M-level attribute(s). Since the number of item parameters increase quickly for this specification, the method assumed that each respondent is either above or below a certain level of mastery. Above the threshold level, an individual is considered to have required mastery to answer the corresponding item(s) correctly. The Q-matrix denotes the level (say  $m$ ) of each skill needed to answer correctly. However, the framing of the problem turns the skill into a dichotomous specification, allowing for the remainder of the model to match a G-DINA.

[150] also proposed a method to incorporate polytomous attributes into the DINA model. As a further extension, the Continuous Conjunctive Model (CCM) proposed by [151], while continuing to assume a dichotomously specified Q-matrix, estimated examinees' mastery of a skill along a continuum. However, this model did not include any item parameters and instead focused on estimating ability profiles based on response patterns.

#### **1.2.2.5.2 IRT-based Cognitive Diagnosis**

When considering latent abilities along a continuum, it is important to trace back the origins of CDM and how they are tied to IRT. [121] used the term IRT-Based Cognitive Diagnostic Models and described a method that was a building block of the current form of CDM. In this IRT-based version, a regression is estimated with item difficulties as the response and the rows of the Q-matrix as the explanatory variables. The overall goal is to estimate a classification of respondents into skill classes, probabilities of occurrence of skill classes, and attribute-specific mastery level. Ability scores from IRT are triangulated with unusual scoring patterns. These scores are estimated separately through fitting IRT models on other large datasets for the given assessments, and not as part of the CDM

estimation.

In the Special Issue 4 of Journal of Educational Measurement (Volume 44), [152] discussed a selection of IRT-based approaches to skill diagnosis for continuous latent traits. All of these models are based on continuous attributes only and do not include any binary or categorical skills. Fischer’s LLTM (Linear Logistic Test Model) [153, 154] modeled a skill  $\theta$  for each individual which can be thought of as a weighted average of proficiency levels on all sub-skills needed to answer an item correctly. The Compensatory Multidimensional IRT (MIRT-C) [155] was essentially a 3 parameter (3PL) IRT with the ability and discrimination parameters being  $K$ -dimensional to reflect an individual’s mastery over / an item’s discriminatory power vis-a-vis each skill needed to respond to an item correctly. The non-compensatory (equivalent of “And” in DINA) version used a 2PL IRT [156] instead of a 3PL in [155]. Similar to the non-compensatory version, the probability of person  $i$  responding to item  $j$  correctly, given that the item needs a subset of the  $K$  total skills, is dependent on the person’s mastery over the given skills and the item’s difficulty and discriminatory power vis-a-vis those skills. The Multicomponent Latent Trait Model (MLTM) [157–159] included an additional term for the probability of a person succeeding on a specific skill needed for a given item. Therefore, there are two multiplicative components in this model. First, a component capturing the possibility that the person can successfully execute all skills needed for an item and can, hence, try to answer the question correctly as a result. The second component captures guessing which has to occur in case the respondent fails on one of the required skills. The probability of success on an individual skill is modeled as a 1PL IRT. Finally, the General Component Latent Trait Model (GLTM) [157, 158] is a combination of an LLTM and an MLTM wherein the top level model is MLTM with the difficulty parameter in the 1PL broken down into sub-sub-skills which contribute to the difficulty of a given skill  $k$ .

### 1.2.2.5.3 Continuous Responses

So far we have discussed CDMs for dichotomously graded assessment responses. However, various situations can lead to continuous responses. For example, survey responses may record endorsement of or agreement with a statement along a continuum which can include a likert scale. Item responses may be graded on a detailed partial scoring scheme. Probability testing frameworks ask respondents to indicate probabilities of each answer choice being the correct one. Finally, time taken to complete an assessment can also be considered a response variable for a CDM, requiring an extension of existing methods. [160] proposed C-DINA (Continuous DINA), an extension of the DINA model for continuous

response. In this framework, response variable follows a lognormal distribution which is considered most appropriate when response is response time. [161] further extended this method to G-DINA, the C-G-DINA.

Recent works have incorporated speed of responding to an assessment into the CDM framework in various ways. [162] proposed a method to incorporate response times into the DINA model to improve classification of skills and skill profiles. [163] incorporated response times in adaptive testing scenarios where the total test time is limited. [164] combined response times with response accuracy to measure fluency in using required skills to answer test questions. [165] discussed a framework for using response times in addition to item responses when modeling random guessing behaviors. Finally, [150] proposed the Joint Differential Speed DINA (JDS-DINA) model to incorporate variable cognitive processing speeds with polytomous attributes to support better feedback and testing strategies.

#### 1.2.2.6 Applications of CDM

The fraction-subtraction dataset introduced by [166] is the canonical application of CDM discussed extensively in the literature. The original dataset includes  $N = 2144$  responses on  $J = 20$  items involving subtractions of fractions. Eight attributes ( $K = 8$ ) were described by [166] and formalized into a Q-matrix by [167]. This dataset or a subset of it has been analyzed to introduce methodological developments of CDM in [122, 128, 141–143, 151, 168, 169], to name a few. [126] observed that the repeated use of two datasets - the fraction-subtraction data (46% in articles found by [170]) and the Examination for the Certificate of Proficiency in English (ECPE) data - is a double-edged sword. The advantage is that this allows for maintaining comparability across methods. However, this may have limited wider applications of the methodology to problems from other areas.

Beyond the use of these traditional datasets, CDM has been applied to other mathematics and language testing situations. The Trends in International Mathematics and Science Study (TIMSS) Assessment data from 1999, 2007, or 2011 have been analyzed by [136, 145, 146, 171, 172]. Other applications to mathematics testing include [147, 173], both of which use data from the Program for International Student Assessment (PISA). Applications of CDM to language or readings tests have been conducted extensively [123, 134, 135, 174–182]. [183] applied CDM to an assessment of orthographic processing of language. All these applications have spanned various geographical areas.



Even though applications of CDM focus extensively on mathematics and language testing, [170] also found papers where the construct of interest pertained to mental health or civic education. It is important to note that [170] only focused on articles in which applied data analysis was the focus of the work. [129] discussed applications of CDM to the measurement of psychological disorders. [184] discussed an application in physics education. There is no commentary in the literature on the complexity involved in developing Q-matrices across disciplines.

#### **1.2.2.6.1 Application to Statistics Education**

Based on our understanding of the literature, an application of cognitive diagnostic modeling to statistics education, and specifically the measurement of statistical literacy, would be a novel contribution. In addition to the disciplinary novelty, there is limited published work discussing cognitive underpinnings of statistical problem solving for a specific assessment instrument. [69] considered the cognitive load of statistical problem solving in the context of transfer distance. [16] outlined the thought process involved in statistical problem-solving, identifying that the synthesis of context and statistical ideas sets it apart from mathematical problem solving. Previous work has mapped cognitive skills to statistics learning under Bloom's taxonomy ([185]; revised [186]). [187] related the statistical literacy, reasoning, and thinking framework to the taxonomy, identifying statistical literacy [8] to be at the lowest level of the cognitive load hierarchy. [188] related stages of their Problem Solving Approach to levels of the taxonomy. [5] argued that statistically literate behavior is an outcome of joint knowledge and dispositional factors, reminding the reader that such behavior requires something beyond cognitive skills. With this view, we argue that the present work makes a contribution to the literature by applying CDM to statistics education and by developing a Q-matrix for statistical literacy measured using research-based assessments.

The remainder of this dissertation is organized as follows. Chapter 2 addresses the first two research questions: (RQ1) Can an isomorphic instrument measure the same underlying construct as the original if all isomorphic items are dependent on relevant contexts? (RQ2) Do students perform comparably on both these assessments? It discusses the development and pilot of the isomorphic assessment instrument designed to measure contextualized statistical literacy. Chapter 3 discusses the design, implementation, and results of the curricular experiment designed to address the following three research questions: (RQ3): does introducing relevant contexts in a statistics classroom cause a differential gain in statistical literacy outcomes? (RQ4): does taking an assessment

of contextualized statistical literacy as a pre-test cause a differential gain in statistical literacy outcomes? (RQ5): does the interaction between contexts incorporated into the classroom and type of statistical literacy assessment cause a differential gain in statistical literacy outcomes? Finally, Chapter 4 discusses the development of a Q-matrix and analyses of data from the pilot study in Chapter 2 using the Cognitive Diagnostic Modeling framework. It answers the following two research questions: (RQ6): Over and above the component skills identified as important for answering questions on the assessment for statistical literacy, does a latent ‘context familiarity’ affect the probability of correctly answering the items? (RQ7): Can the modified assessment of contextualized statistical literacy (MBLIS) provide feedback on the same statistical skills as BLIS?

# Chapter 2 | Examining the Role of Context in Statistical Literacy Outcomes using an Isomorphic Assessment Instrument<sup>1</sup>

## 2.1 Introduction

The importance and role of statistical literacy has been discussed extensively in the statistics education literature [1–14]. Guiding documents which inform researchers and practitioners alike, such as the GAISE College Report [18], Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) Research Report [19], International Handbook of Research in Statistics Education [1], and GAISE PreK-12 Report [20], highlight this importance vis-a-vis cognitive outcomes, curriculum, teaching practices, and assessments. The American Statistical Association discusses ‘(to) build a statistically literate society’ as one of its objectives under the strategic goal of statistics education [21]. In parallel, the PARIS21 partnership [22] among global organizations including United Nations, European Union, Organisation for Economic Co-operation and Development, International Monetary Fund, and the World Bank also considers statistical literacy to be a focus of its work. Even though definitions of statistical literacy vary in some aspects [23], mainstream conceptualizations of statistical literacy agree that it comprises of a skillset which an individual would benefit from applying to contexts outside of a classroom, a skillset which would allow people to engage with contexts

---

<sup>1</sup>This chapter, including the associated Appendix B, has been submitted as an article to the Statistics Education Research Journal (SERJ).

relevant to them from a data-driven point of view. Further, the literature also converges on a firm belief that statistical literacy plays a critical role in promoting a citizenry that is more capable of understanding the world around them and making evidence-based decisions in their private and public lives. Under this premise, our work asks the following question: Are students of statistics able to make sense of statistical insights encountered in their day-to-day lives, especially pertaining to relevant topics? Such an ability is considered to be an important marker of a statistically literate citizen [24].

### **2.1.1 Role of Context and Transfer**

Considerable amount of work has discussed the value of contexts and powerful ways of introducing contexts which are familiar to the students into the curriculum [26, 44–47]. Concurrently, studies focusing on improving statistical literacy among students at various levels have also been conducted [27–32]. However, even though previous work has measured statistically literate behavior outside of a classroom setting [60, 61], there is limited work proposing research-based assessments of statistical literacy [62–64]. Assuming that applying statistical literacy skills to new contexts would involve a knowledge transfer [65–67], we distinguish statistical literacy skills from the ability to apply those skills to topics relevant in our lives, and define contextualized statistical literacy. Such a transfer, though it is central to the purpose of statistical literacy, is not encoded in the definition of statistical literacy. Contextualized statistical literacy is statistical literacy as it pertains to relevant contexts, where relevant contexts are conceptualized as ones that are societally relevant at a given time and people would have engaged with and thought about on their own. The key contribution of this work is in creating an instrument to measure contextualized statistical literacy using an existing research-based assessment of statistical literacy allowing us to examine respondents' statistical literacy skills when they are required to apply those in relevant contexts.

As underscored by [68], the terms near and far transfer are relative and the distance of transfer implied in those terms is open to interpretation. [69] discusses distance vis-a-vis the similarity to problems encountered during instruction. [70] highlight that distance of transfer is an intuitive notion and discuss it as a matter of similarity and familiarity. Near transfer is across contexts students can be expected to be familiar with because they have encountered similar contexts before during instruction or practice. Whereas far transfer involves transfer across contexts which may not be similar, on the surface, to anything students have encountered before. According to [67], any application of statistical literacy skills can be considered to be a transfer problem. However, when

considering contextualized statistical literacy, the question of distance of transfer is not straightforward. On one hand, encountering statistical constructs in new contexts increases the distance of transfer. On the other hand, though, irrespective of whether or not these relevant contexts have been introduced in the classroom before, since we conceptualize relevance as familiarity and engagement outside of the classroom, it can be considered to be nearer transfer for a respondent of an assessment of contextualized statistical literacy. When considering this transfer, we must also be also mindful of possible suspension of sense-making [71, 72] whereby familiarity with the context maybe foregone in favor of focusing on the underlying statistical idea.

### **2.1.2 Isomorphic Assessment**

To measure the transfer of statistical literacy skills to relevant contexts, we created an isomorphic version of an existing research-based assessment of statistical literacy. An isomorphic question or item is identical to a base item in structure (concept, phrasing, as well as distractors) and differs only in the context, continuing to measure the same underlying construct [73–75]. Isomorphic items can also be visualized as items with a common base template differing only in context [68]. [76] and [77] refer to these as structural isomorphs to highlight that this framework itself does not guarantee that respondents’ cognitive processes in answering these tasks will be comparable. Isomorphic tasks have been studied extensively in the physics education literature [78–83]. Some work has also been conducted in the computer science education domain [74, 84]. It is worth noting that there is limited work in the statistics education research literature which studies isomorphs. Most of the aforementioned studies deployed isomorphs to gauge learning and understand common misconceptions, and designed the study in such a way that each respondent solved all of the two or more isomorphic problems at different time points in a random order. [78] study was the only exception where each respondent was assigned to one of the two versions of the assessment. Previous work using isomorphic tasks finds that transfer across such tasks is difficult in most circumstances even if only incidental features are switched. There is some evidence, e.g. [79], that more practice on the base topics improves performance as discussed by [67]. The findings in [85] are valuable given the objective of this work. Their work studied transfer across contexts which cross the disciplinary boundaries in which the construct is situated. They studied the effects of algebraic training on contexts within mathematics as well as in physics to find that training in mathematics facilitated transfer to physics but not the other way round. Even though every context in a statistics problem is external to the discipline

itself, this work is important to consider because it provides some evidence of facilitating transfer where contextual information has been provided. This would indicate that familiarity with relevant contexts should improve student performance on statistical literacy tasks as compared to an isomorph based on potentially unfamiliar tasks barring any suspension of sense making [71, 72].

These studies of transfer using isomorphic tasks deploy a variety of types of assessments. However, very few of them use research-based assessments. [84] discuss the importance of developing assessment instruments which undergo rigorous process of collecting reliability and validity evidence, and for researchers to adopt these for further research. [86], in outlining ‘a practical approach to validation’ of research-based assessments support the value and importance of this in step 4 - ‘Identify candidate instruments and/or create/adapt a new instrument’ - with a reminder to first look for previously developed instruments. We chose the Basic Literacy in Statistics (BLIS) assessment [17, 64] because of it’s sole focus on statistical literacy and the extensive research conducted to gather reliability evidence and develop a validity argument for its intended use. We created an isomorphic version of BLIS, i.e., M-BLIS hereafter, to answer the following research questions: (RQ1) Can an isomorphic instrument measure the same underlying construct as the original if all isomorphic items are dependent on relevant contexts? (RQ2) Do students perform comparably on both these assessments? This chapter describes the development of M-BLIS, and the design and analysis of a study comparing the original and the isomorphic assessment instruments.

## 2.2 Methodology

This section discusses the development of M-BLIS, design of the pilot study, and statistical methods used to analyze data from the study. Section 2.2.1 details the process of creating M-BLIS, including the choice of relevant contexts and parameters considered when developing isomorphic items. Additionally, it provides examples of modified tasks. We then outline (Section 2.2.2) the expert review process and the pilot study conducted at a large public research university in the eastern United States. Data from the pilot study are analyzed with the two research questions (RQs) outlined above. Finally, Section 2.2.3 describes the analytical methods.

### 2.2.1 Assessment Modification

Once BLIS was chosen as the instrument for measuring statistical literacy, it was modified to include isomorphic items. These isomorphic items were intended to depend on ‘relevant contexts’ - contexts which are societally relevant at the time and test-takers would have engaged with on their own outside of, and apart from, class. Various topic options such as climate change, immigration, race-related issues, and the COVID-19 pandemic were considered. The pandemic impacted the life of all individuals, presenting a unique opportunity for research in the form of a topic everyone could engage with and find relevant. We acknowledge the devastating effects of the pandemic and the psychological impact this may have on individual test-taker’s performance. If a test-taker’s performance is adversely affected by their emotional reaction to a given context (e.g., due to traumatic hardship or death of a loved one), we have limited ability to separate out the effect of the emotional reaction from their conceptual understanding. We must also consider the ethics of compelling a test-taker to look at statistics about a potentially sensitive topic such as the number of deaths due to the pandemic. Having said that, this issue is not unique to the COVID-19 pandemic. There is a broader question of whether and how to incorporate potentially sensitive topics on assessments while balancing the competing goals of encouraging students to look at relevant societal issues from a statistical lens on one hand, and the ethical and measurement-related issues arising as a result of the sensitive topic affecting test-takers differentially on the other hand. For the purpose of this work, we decided to proceed with the COVID-19 pandemic as the broader context with a clear intention of excluding any statistics pertaining to severe illness, loss of life or livelihood, and other serious health effects including but not limited to mental health. While mortality and hospitalization during the COVID-19 pandemic were almost certainly engaging contexts, the strong negative association for students who may have endured such trauma or loss would likely evoke an emotional response. Such contexts were avoided mainly as a matter of compassion. Additionally, though, such contexts could have prevented the respondent from completing the assessment to the best of their ability. Two competing requirements of improving engagement and reducing emotional impact were balanced during the process of creating the isomorphic assessment. Contexts such as sleep length among college students, dental care choices, flight cancellations, restaurant visit frequency, and air pollution, all during the COVID-19 pandemic, were incorporated.

Each item went through extensive considerations. Before looking for data and reports from which contexts and statistics were sourced, following were noted for each item

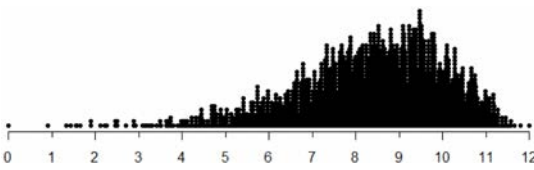
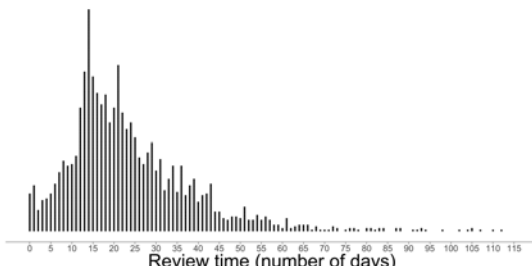
on BLIS: type(s) of variable(s), parameter(s) of interest, type of sample, type of study (observational versus experimental), and whether creation of the item required access to raw data or summary statistics or neither. These considerations are presented in the modified test blueprint (Table A.2). Additionally, keywords were considered in an attempt to find a context which could match the original item as closely as possible. Table 2.1 demonstrates an item for which an alternative study with a highly comparable context was found. The original item was based on a real but not widely relevant context. This item may be considered one of the purest forms of an isomorph in M-BLIS. The **bolded text** highlights common words across the two versions, excluding singulars or plurals.

Original item stem	Modified item stem
<p><b>Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up different scents. An experiment was conducted with a dog in Japan who was trained to smell bowel cancer in stool samples. In a test, the dog was presented with five stool samples; one from a cancer patient and four from healthy people. The dog indicated which stool sample was from the cancer patient. This was repeated a total of 38 times. Out of the 38 tests, the dog correctly identified the cancer sample 37 times. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies cancer more than one fifth of the time. The p-value is less than .001. Assuming it was a well-designed study, use a significance level of .05 to make a decision.</b></p>	<p><b>Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up different scents. A pilot experiment was conducted with dogs in Germany who were trained to smell COVID-19 in saliva samples. In the test, one dog was presented with 115 saliva samples; 21 from COVID-19 patients and 94 from healthy people. The dog indicated which saliva samples were from the COVID-19 patients. Out of the 21 COVID-19 positive samples, the dog correctly identified 20 of them. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies COVID-19 more than half the times. The p-value is less than .001. Assuming it was a well-designed study, use a significance level of .05 to make a decision.</b> Source: Research article.</p>

Table 2.1. Real from real data to Real from relevant data



Each original item was also categorized based on the scale discussed in the GAISE College Report [18]: Naked data, Realistic data, Real data, Real data from a real study. This is reflected in the table caption. This categorization served a dual purpose. The first and broader purpose of considering this categorization was to analyze whether any observable effect is associated with with the degree of change from the original data category to the modified category (real data from a relevant study). Secondly, it allowed us to understand which isomorphs needed to go beyond simply replacing context-specific words. For example, the item in Table 2.2 was based on naked data in the original BLIS. However, given the purpose of this work, the change had to go beyond a simple isomorph.

Original item	Modified item
<p data-bbox="240 231 803 304">The distribution for a population of measurements is presented below.</p>  <p data-bbox="240 871 803 1092">A sample of 10 randomly selected values will be taken from the population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?</p> <ul data-bbox="272 1144 414 1365" style="list-style-type: none"> <li>• 6 to 7</li> <li>• 8 to 9</li> <li>• 9 to 10</li> <li>• 10 to 11</li> </ul>	<p data-bbox="836 231 1477 598">For scientific credibility, journal articles are reviewed by other scientists before publication. This process is called peer-review. Researchers collected data to study how the pandemic has affected the peer-review timelines for six Ecology journals. The plot below shows the distribution of number of days taken by all reviewers to review papers assigned to them.</p>  <p data-bbox="836 871 1477 1092">A sample of 10 randomly selected papers will be taken from this population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean? Source: Research article.</p> <ul data-bbox="868 1144 1015 1365" style="list-style-type: none"> <li>• 0 to 10</li> <li>• 10 to 20</li> <li>• 20 to 30</li> <li>• 40 to 50</li> </ul>

**Table 2.2.** Naked to Relevant

In contrast, Table 2.3 demonstrates an item which was based on real data from a real study, leading to an isomorph that retained the structure of the original item very closely.

Original item stem	Modified item stem
<p>The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population.</p>	<p>The Pew Research Center surveyed a nationally representative group of 12,648 U.S. adults in November 2020. Of these adults, 62% said they would be uncomfortable being among the first to get the vaccine for COVID-19. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population. Source: Pew Research Center.</p>

**Table 2.3.** Real from real data to Real from relevant data

For each modified item, the source link was provided at the end of the prompt. It was added on a separate line with the word “Source” followed by a very short key phrase identifying the source with a hyperlink. This was intended to underscore the authenticity and credibility of the contexts presented in the item without distracting the test-taker from the key task. During a think-aloud conducted prior to field testing, a respondent explicitly stated that this added legitimacy to the questions in the student’s mind.

The stringent requirement to retain the structural integrity of item phrasing and the statistical idea in the isomorph was loosened for two items. The modified context for the item in Table 2.4 was powerful enough to compel such a concession.

Original item stem	Modified item stem
<p>According to the National Cancer Institute, the probability of a man in the United States developing prostate cancer at some point during his lifetime is .15. What does the statistic, .15, mean in the context of this report from the National Cancer Institute?</p>	<p>Consider an individual fitting the following description.</p> <ul style="list-style-type: none"> <li>• 20-year-old female,</li> <li>• lives alone near a university campus,</li> <li>• is exposed to an average of 10 people each week,</li> <li>• has no underlying medical complications,</li> <li>• is asymptomatic and unvaccinated,</li> <li>• and follows CDC's guidance.</li> </ul> <p>According to the "19andMe" tool developed by Mathematica, her probability of catching COVID-19 through community transmission in a week is .0024, as of March 30, 2021. What does the statistic, .0024, mean in the context of this calculation from Mathematica? Source: Online calculator.</p>

**Table 2.4.** Context changed considerably

Finally, the item in Table 2.5 was a subject of lengthy discussions, some of which included the expert reviewers. The implicit assumption of a coin being unbiased and our intuition about 50% of them landing on heads benefitted the original item. However, upon deliberation, it was agreed that it is extremely hard to find other phenomena which have an unconditional 0.5 probability of occurrence which is understood intuitively, and therefore the substantial change in wording was included.

Original item stem	Modified item stem
Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 50 times and the other student flips a coin 100 times. Which student is more likely to get 48% to 52% of their coin flips heads up?	Penn State University administrators surveyed all undergraduate students to capture feedback from the entire student body on several issues. As a result, they learned that 86% of all students planned to return in fall 2020. Despite knowing the proportion for all Penn State students as a whole, several instructors surveyed their own classes in order to be sensitive to the views of their students. One instructor had a class with 50 students and another instructor had a class with 100 students. Assuming both classes were representative of the entire student body at Penn State, which instructor was more likely to find that 84% to 88% of their students would plan to return in fall 2020? Source: Adapted from Penn State News.

**Table 2.5.** Implicit assumption changed

In addition to the factors discussed above, we also considered peculiarities such as the distance of the sample statistic from the parameter, scale of the p-value, whether a small sample was required, and overall length (in characters) of the names of variables or context description. The original assessment was unchanged from the version provided in [17]. For items that involved a visualization, plots were created using the *ggplot2* package [189] in *R* [190]. Even though some of the original visualizations were created using a the *plotrix* package [191], same aesthetics and scales were maintained in the modified visualizations.

### 2.2.2 Study Design

To investigate whether M-BLIS continues to measure the same underlying constructs and whether students perform comparably on both assessments, we gathered data to generate reliability evidence and develop a validity argument using methods recommended in the *Standards* [192]. These included expert reviews, think aloud interviews, and a pilot study, though they were adapted to be suitable for the development of an isomorphic assessment instead of a new instrument.

Three expert reviewers looked at the modified instrument with the prompt, “Please

consider each modified item vis-a-vis the original item and comment on whether they are comparable in measuring the underlying learning outcome.” The instrument was updated based on expert feedback for a pilot study. In the pilot instrument, six out of the 37 total items were retained as anchors. This allowed us to equate scores under the internal-anchor design discussed in [193].

This updated instrument was piloted in a study conducted at a large public research university in the eastern United States. The pilot was designed to address the two questions described above. (RQ1) How does the functioning of the isomorphic items compare to the functioning of the original assessment? (RQ2) Is there evidence to suggest that test-takers respond to the underlying statistical question differentially if the item is based on a relevant context? To facilitate this investigation, we built two levels of comparisons. First, consenting test-takers were randomly assigned to take either BLIS or MBLIS. This gave us a baseline on the original assessment within the target population, facilitating a comparison of results across the results from [17]’s field test and our pilot study. This allowed us to answer the main research question - is MBLIS measuring the same constructs as BLIS? To add another layer of comparison, five randomly chosen items from the original assessment were retained as anchors. One of the randomly selected item numbers was a part of a testlet leading to six anchor items. Resultantly, M-BLIS featured 6 original and 31 modified items. Alternative criteria such as model-based difficulty ranking and topic were considered for the determination of the anchor items. However, a random selection was decided to be the best choice at the end.

The original instrument was developed specifically for an undergraduate introductory statistics student learning under the simulation-based inference curriculum. Although learning under a different framework (Lock5 curriculum instead of CATALST), the undergraduate introductory statistics course at the aforementioned university matched this description, providing in its students an easy choice of group to use as pilot subjects. The original instrument was studied as a mid-semester and end-of-semester evaluation. Therefore, the pilot study deployed it as a post test in the aforementioned class.

Finally, an extensive survey was attached at the end of the assessment to learn student demographics, their interest in and engagement with certain relevant topics such as diversity questions, immigration, politics and governance, and their experience of interacting with items pertaining to the pandemic (given only to M-BLIS respondents). The last subset included questions such as whether responding to items regarding the pandemic was discomforting to them and whether their ability to consider the statistical question was affected by the contexts.

The data collected from this pilot will be used to further validate M-BLIS. Learning about differential performance across the two assessments within similar subgroups of students will provide us with useful information on the performance of the modified instrument. Further, item-by-item comparison across the instruments will allow us to look more closely at the items which may perform differently. These findings will be most instrumental in us developing the validity and reliability argument for M-BLIS.

### 2.2.3 Statistical Methodology

Data from the pilot study were analyzed with two key goals in mind. First, are the two assessments (BLIS and M-BLIS) comparable in measuring underlying constructs? Second, did the test-takers perform comparably on both the assessments? Though both sets of analyses were conducted using the same student performance data, separating out these two goals helps us discuss results accordingly.

To compare the two assessments themselves, we evaluated measures of reliability and measures which can contribute towards a validity argument for the use of the instrument. All the measures in this set are replicated based on the analyses in [17]. As a measure of reliability, we consider the coefficient of alpha to compare internal consistency among items. In order to check the assumptions of Item Response Theory (IRT) models which contribute towards the validity argument, we conduct principal component analysis to check two assumptions of the IRT models - unidimensionality and local independence. Scree plots based on PCA allow us to comment on that. Single-factor confirmatory factor analysis allow us to further comment on unidimensionality. Finally, we also fit Rasch, 2PL, and PC models and look at item information curves from the best of these models to learn whether item difficulties are comparable across the two assessments. Test information functions and standard errors of measurement are also considered for each instrument. On the validity side, we look at item parameters for partial credit (PC) model with 32 items and 4 testlets, and item characteristic curves for each item or testlet. It is important to note that even though we may occasionally compare our results to the original study conducted in form of field test in [17], a separate set of analyses where the pilot study is considered a replication of the original study will be discussed in forthcoming work. For the purpose of this discussion, we limit ourselves to analyses which consider whether the two instruments, as suggested by data from our pilot, function comparably with each other in terms of the reliability measures and pieces of the validity argument.

To compare student performance more directly, we analyzed data under the classical

test theory (CTT) framework. Though we acknowledge the advantages of using the IRT framework when analyzing assessment data, CTT was preferred due to the ease with which the relationship between test-taker covariates and performance can be interpreted in the models. Even though recent developments have introduced IRT models with covariates, CTT-based models are also easier to interpret. We fit multiple linear regression models - with total scores as the response variable - to understand the differential performance across two assessments. Type of assessment randomly assigned to a student was the key explanatory variable of interest. We also included student demographics such as their gender, whether they are an international student or not, and highest education of a parent/guardian, as well as responses to pertinent survey questions. These analyses were conducted using all responses, as well as the complete-only responses. In this paper, we present the latter.

## 2.3 Results

### 2.3.1 Comparing Assessment Instruments

In this section, we address the first research question (RQ1): do the two instruments, BLIS and M-BLIS, perform comparably?

#### 2.3.1.1 Summary of Assessment Performance

First we summarize item-by-item performance to highlight key differences. Table 2.6 contains percentage of correct responses per item. The Difference column is calculated as  $\text{correct}_{\text{BLIS}} - \text{correct}_{\text{M-BLIS}}$ . Table B.7 in the Appendix B tabulates selected-responses i.e. percentages of respondents who chose each distractor. Shaded blue rows indicate anchor items which are critical in comparing the two groups of respondents at baseline. Shaded orange rows highlight items with difference values around or higher than an arbitrary cutoff of 10%.

Item	BLIS	M-BLIS	Difference	BLIS context - GAISE
Q1	74.6	73.2	1.4	Real from real study
Q2	44.0	50.7	-6.7	Realistic
Q3	53.4	52.5	0.9	Real
Q4	83.5	86.2	-2.7	Real
Q5	81.3	84.7	-3.4	Realistic



Item	BLIS	M-BLIS	Difference	BLIS context - GAISE
Q6	73.5	70.7	2.8	Realistic
Q7	35.6	41.1	-5.5	Real from real study
Q8	29.5	32.8	-3.3	Realistic
Q9	65.4	34.0	31.4	Realistic
Q10	56.3	39.2	17.1	Realistic
Q11	42.0	37.1	4.9	Real
Q12	58.3	48.8	9.5	Real from real study
Q13*	37.6	37.6	0.0	Real from real study
Q14	42.8	24.6	18.2	Naked
Q15	63.8	48.5	15.3	Realistic
Q16*	24.6	27.8	-3.2	Realistic
Q17*	46.1	46.8	-0.7	Realistic
Q18	45.9	45.4	0.5	Real from real study
Q19	40.8	38.4	2.4	Real from real study
Q20	37.9	34.3	3.6	Real from real study
Q21	16.5	16.3	0.2	Real from real study
Q22	58.5	61.0	-2.5	Realistic
Q23*	43.4	43.9	-0.5	Real from real study
Q24*	57.2	60.0	-2.8	Real from real study
Q25	55.5	61.6	-6.1	Real from real study
Q26	42.2	42.0	0.2	Real from real study
Q27	38.6	45.0	-6.4	Realistic
Q28	52.7	60.2	-7.5	Real from real study
Q29	52.0	48.9	3.1	Real from real study
Q30	48.3	45.5	2.8	Real from real study
Q31	86.4	83.9	2.5	Realistic
Q32*	48.0	43.6	4.4	Real from real study
Q33	64.4	62.0	2.4	Real from real study
Q34	70.4	65.2	5.2	Real
Q35	23.4	21.6	1.8	Real from real study
Q36	79.0	68.6	10.4	Real
Q37	57.8	54.3	3.5	Real from real study

Item	BLIS	M-BLIS	Difference	BLIS context - GAISE
------	------	--------	------------	----------------------

Table 2.6: Difference in proportion of respondents correctly answering each item

**Anchor items:** All anchor items (13, 16, 17, 23, 24, 32) had a difference of less than 5% in the proportion of respondents who chose the correct answer. On four of the five remaining items, two of which formed a testlet (*items 23-24*), the M-BLIS group had a higher percentage of respondents choosing the correct answer. On *item 32*, students in the BLIS group performed better. However, authors must note here that there was an inconsistency in the presentation of *item 16* across the two versions. Aside from this *item 16*, the distribution of selected responses was comparable across the two assessments, providing evidence of the two randomized groups being comparable at baseline.

**Isomorphic items:** Excluding the anchor items, 22 out of the remaining 31 items witnessed better performance on BLIS. Further, the items with the highest absolute difference were ones where the BLIS saw better performance. The items where students performed better on M-BLIS are unevenly distributed in terms of topic coverage. The latter half of the assessment was based on inferential statistics and it appears as though BLIS was easier for those topics. This gives us an early indication that items with relevant context may have been more difficult to answer. Having said that, about half the differences were less than 5%, suggesting that the instruments may be more comparable than suggested by the extreme difference values.

Six items had absolute differences close to or higher than 10%. Respondents performed better on BLIS on all of these items. Five out of these six items pertained to graphs and descriptive statistics, and the sixth was based on regression and correlation. The difficulty levels of these items, according to [17]’s analysis, were well-distributed. Two of these six items were discussed in Section 2.2.1. The BLIS item which was naked (Table 2.2) and the item where length was changed considerably to accommodate a pertinent relevant context (Table 2.4) were both in this group. Further, the item which changed from naked to relevant saw an incorrect option being chosen more frequently than the correct one, even though this observation must be treated with care since the direction of the skewness was reversed.

Item 10 warrants a closer look not only because of the high difference in proportion, but also because of the contexts. The BLIS context for this item was ‘number of hours of sleep for college students,’ whereas the M-BLIS context was ‘an index capturing

the strictness of lockdown policies implemented by various countries’ governments in response to the COVID-19 pandemic.’ This is also true, with a higher difference value, for item 9. The BLIS context was ‘self-reported confidence about success of students in an introductory statistics class,’ and the M-BLIS context was ‘rating from Vietnamese citizens indicating how overwhelming the official news regarding the pandemic has been for them.’ In contrast, though, four other items (5, 6, 8, and 35) also featured contexts specific to college students on BLIS and the differences are much smaller for those items.

Finally, we looked at two specific items of interest. Of the five testlets included in this instrument, only one shared a learning outcome - items 29 and 30. Item 29 required a respondent to choose the correct null hypothesis and 30 required them to select the alternative hypothesis for the same stem. Table 2.7 tabulates performance on these two items.

	BLIS		M-BLIS	
	Correct_30	Incorrect_30	Correct_30	Incorrect_30
Correct_Q29	40.9	11.1	35.1	13.8
Incorrect_Q29	7.4	40.6	10.4	40.7

Table 2.7: Performance on testlet items

In the original study, the author decided to treat items 29 and 30 as a testlet because majority of the students either got both questions right or both of them wrong. Since this observation holds true for M-BLIS as well, we treat our data as a 36-item-scale (with one testlet) for the remaining analyses.

### 2.3.1.2 Reliability Evidence and Evidence for Validity Argument

The expert reviews contribute to the evidence used to develop a validity argument for the intended use of M-BLIS. A large proportion of expert comments entailed a change in the structure of the original instrument itself. These comments ranged from minor changes in wording to discussions regarding whether the item is measuring the construct it claims to. However, any comment which would have led to a change in the original item was marked as out of scope for the purpose of this work. Of the 31 items (six anchors) under consideration, 12 were unchanged and 16 received minor wording and presentation changes. The language and presentation of one item and all its distractors was updated significantly. Two items were topics of lengthy discussions. They were almost completely

rewritten based on expert reviews. One of these two items is presented in Table 2.5 and discussed subsequently. The other isomorph that underwent a significant change based on expert reviews shared a key characteristic with the item in Table 2.5. The context in that item also had a binary outcome which could be intuitively assumed to be equiprobable. In both these cases, the original context of our choice was retained while rephrasing the item stem. The resulting instrument based on these changes was deployed in the pilot study discussed in Section 2.2.2.

The coefficient alpha indicates internal consistency among test items. In Table 2.8,  $\alpha_{36}$  is the raw coefficient alpha with the single testlet,  $\text{predicted}_{32}$  is the predicted coefficient alpha for the 32-item instrument using the Spearman-Brown formula, and  $\alpha_{32}$  is the actual coefficient alpha for scores with the other 4 testlets. Scores for the other four testlet (excluding items 29-30 discussed in Table 2.7) include partial scoring. The small differences between the 36 and 32 item scales, as well as across the predicted and calculated values for the 32-item scale, indicate that local independence is not a concern among testlet items. These metrics were chosen in order to compare our results with those discussed in [17].

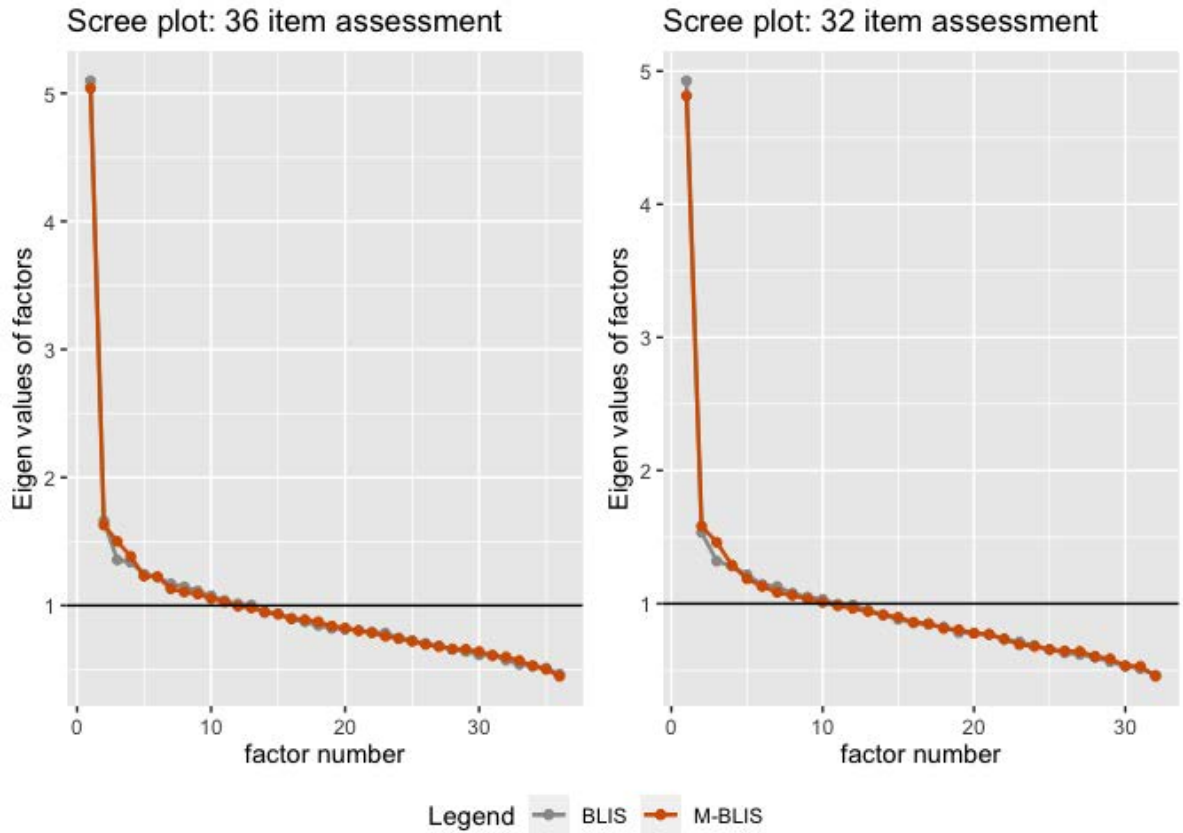
	BLIS	M-BLIS
alpha_36	0.78	0.77
predicted_32	0.76	0.75
alpha_32	0.78	0.76

Table 2.8: Raw and predicted coefficient alpha values

IRT analyses conducted to develop a validity argument make assumptions of unidimensionality in the assessment scale and local independence among items. Results of principle Component Analysis (PCA) are summarized as scree plots in Figure 2.1. The left panel displays scree plots for the 36-item scale and the right panel displays equivalent plots for the 32-item scale with 4 testlets. The two plots within each panel are very similar, effectively hiding the grey dots for BLIS. However, this is encouraging evidence indicating that the two assessment versions are performing comparably.

The eigenvalues level-off after the first factor providing support to the hypothesis that the assessment instruments both consist of a single factor. We do not observe any clear differences between the 36 and 32 item assessments. All the scree plots show evidence of unidimensionality in the instruments. Acceptability of the local independence assumption

was checked using single-factor confirmatory analyses. Results indicated that including testlet scores is acceptable to meet the local independence assumption.



**Figure 2.1.** Scree plots of eigenvalues from PCA

Three IRT models were fit to these data - Rasch model, a 2 parameter logistic (2PL) model, and a partial credit (PC) model. The Rasch and 2PL models were based on 36-item scale, whereas the PC model was based on the 32-item scale incorporating partial grading on the four testlets. Table 2.9 summarizes three model fit measures i.e. the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the Log Likelihood (LL). Based on these indices, the PC model seems to perform the best on both assessments, though the differences in values are fairly small. Therefore, the PC model will be used for the remainder of the analyses.

	BLIS			M-BLIS		
	Rasch	2PL	PC	Rasch	2PL	PC
AIC	27771.08	27171.91	26358.63	26902.39	26246.05	25607.35

	BLIS			M-BLIS		
	Rasch	2PL	PC	Rasch	2PL	PC
BIC	27936.04	27492.91	26523.58	27065.99	26564.41	25770.95
LL	-13848.54	-13513.95	-13142.31	-13414.20	-13051.03	-12766.67

Table 2.9: Model summaries for IRT models

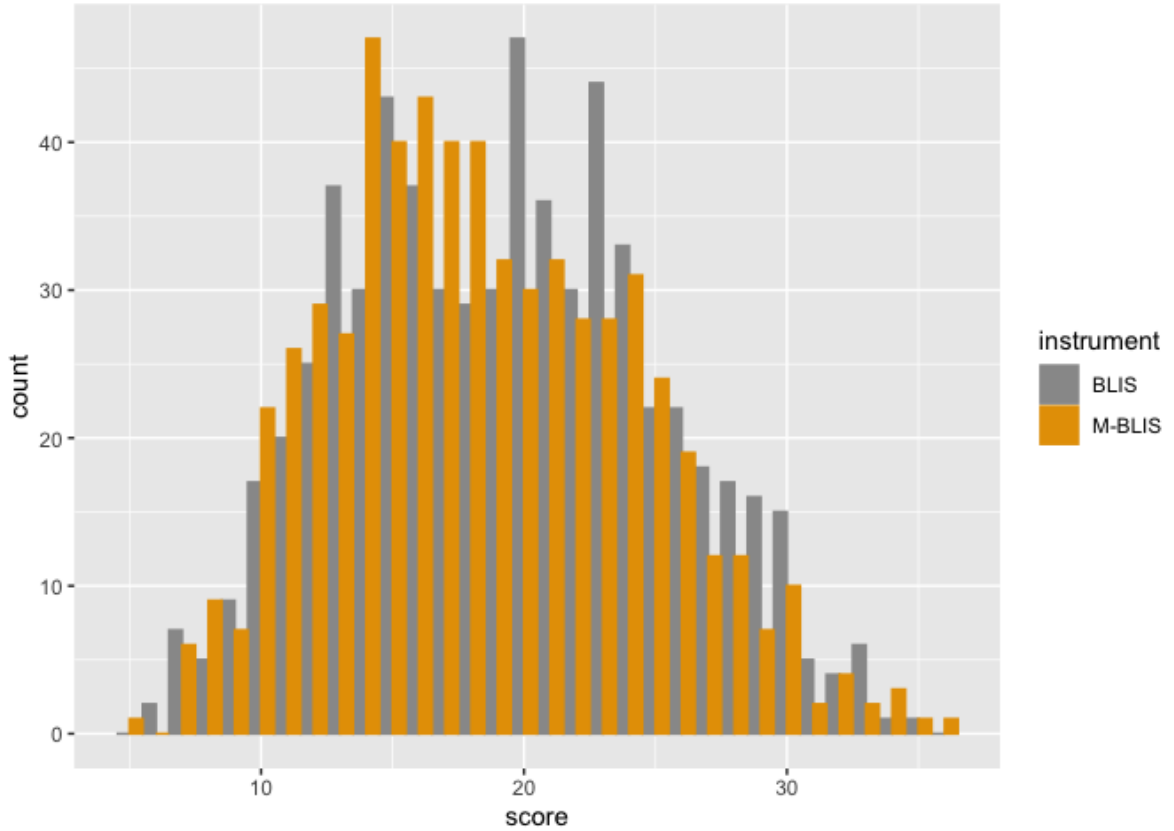
Figure B.2 and Figure B.3 in Appendix B display item information curves. These curves indicate that the instruments contain items which give us information across the ability levels, thereby differentiating test-takers across all ability levels. When comparing the two assessments we notice that there are a few more items on the modified instrument measuring students at higher ability levels than those on the original scale. The test information function and standard error (SE) of measurement curves in Figure B.4 and Figure B.4 (Appendix B) support this observation. Overall test information for M-BLIS is highest at a slightly higher ability level than it is for the original instrument. For BLIS, the SE is slightly lower at lower ability levels, giving slightly more information at lower abilities. For M-BLIS, SE is comparable at the extremes, indicating that it is giving equally little information at the highest or lowest abilities.

Finally, difficulty rankings based on the PC model and item characteristic curves are considered for validity evidence. Table B.8 in Appendix B displays difficulty estimates based on the PC model. They range from -2.02 to 1.78 for BLIS and from -1.98 to 1.78 for M-BLIS. Even though both assessments display comparable ranges of difficulty spread evenly on either both side of zero, the distribution of difficulties is slightly uneven on either side of zero. Half of the 32 items/testlets on the original instrument have difficulty estimates lower than zero. This division is 14 under zero and 18 above for the modified instrument. In line with earlier results, the five items which are most distant in terms of difficulty rankings when BLIS and M-BLIS are compared, are the same items which had higher than 10% difference in Table 2.6. Further, the two items with highest difficulty (consistent across the two instruments) are the two items for which respondents chose an incorrect option most frequently according to Table B.7 in Appendix B. These findings are also supported in the item characteristic curves seen in Figure B.6 and Figure B.7 also in the Appendix B. However, these items have negative correlations with the total score without accounting for the given item. This reverse discrimination indicates a possible flaw in the item and warrants further investigation.

### **2.3.2 Comparing Student Performance**

(RQ2): did respondents perform comparably on the two assessment instruments BLIS and M-BLIS? In this section, we address the second research question by exploring the relationship between assessment performance and assessment type. We also incorporate various test-taker attributes to further understand this relationship. Figure 2.2 shows a distribution of total score by assessment. Assessments are scored as one point per correctly answered question with a highest possible total of 37 points. The grey bars

represent scores on the M-BLIS and yellow bars represent scores on BLIS.



**Figure 2.2.** Comparison of total score (out of 37)

More students scored higher on BLIS. Figure B.1 in the Appendix B plots the two histograms in separate panels, highlighting the slight right skew in the scores on M-BLIS. Overall scores were comparable on both assessments, as suggested by the numerical summaries in Table 2.10. However, a two-sided t-test for the difference in mean scores led to a p-value of 0.005 indicating a significantly lower score on M-BLIS.

Instrument	n	Mean	Median	SD	IQR
BLIS	638	19.31	19	5.98	8.75
M-BLIS	615	18.38	18	5.78	9.00

Table 2.10: Summary statistics of total scores

Before looking at results from statistical models which consider the difference between



BLIS and M-BLIS scores accounting for covariates, univariate summaries of important variables were considered to identify any categories with a small n (Table B.1 - Table B.6 in Appendix B). For all tables, non-responses were removed. There were less than (1%) missing values in any of the variables. Some categories of the gender variable and the group of students who expected to get an F in the class were the only groups with less than (1%) frequency. Bivariate frequencies of all these variables crossed with assessment type were also considered. They ascertained that these variables have comparable distributions across the BLIS and M-BLIS groups.

Four survey questions were considered important for the statistical models in an effort to explore whether they were related to assessment performance. A set of survey questions explicitly asked the respondent about their interaction with COVID-related contexts as well as whether the contexts interacted with their assessment-taking experience. Before including the survey responses into an inferential model, we summarize them. Table 2.11 tabulates percentage responses to three statements “I have actively looked for information on this topic (COVID-19) in the last 6 months.” (**Engagement**), “I would like to gain data-driven insights into this topic (COVID-19).” (**Statistical interest**), “I think this topic (COVID-19) is relevant to our lives” (**Relevance**), each with “Yes”, “Maybe”, or “No” alternatives, based on the assessment type.

Instrument	Engagement			Statistical interest			Relevance		
	Yes	Maybe	No	Yes	Maybe	No	Yes	Maybe	No
BLIS	0.84	0.07	0.08	0.63	0.19	0.18	0.93	0.05	0.02
M-BLIS	0.84	0.07	0.09	0.58	0.21	0.21	0.94	0.03	0.03

Table 2.11: Survey questions regarding COVID-19 pandemic

Frequencies of responses are well-distributed across the two assessments on questions pertaining to the COVID-19 pandemic. The interest question (middle portion of the table) is the only one where some difference in proportion is observed for those selecting Yes, when separated by assessment type. However, the difference looks small enough.

Table 2.12 shows responses to the question “Did the context affect your ability to answer the statistical question?” This question will also be included in the models.

Context affecting ability	Frequency
No difference	0.471

Context affecting ability	Frequency
Made it easier	0.495
Made it harder	0.034

Table 2.12: Self-reported effect of context on ability to respond to the statistical question

Finally, linear regression models were fit to explain the relationship between total scores and type of assessment. Additional variables were included to understand how the effect of assessment type on total score changed in the presence of other test-taker characteristics and their survey responses. This classical test theory-based approach towards analyzing assessment data was preferred due to its focus on the explanatory variables. Following a test for difference in means, various additive as well as interactive linear regression models incorporating a subset or all of the variables tabulated in Appendix B were fit. The adjusted  $R^2$  for all models was in the proximity of 20%. We believe that for assessment data collected in an educational setting, this explanatory power is typical given the plethora of sources of variation. The final model discussed in this paper is per Equation 2.1 and the results presented in Table B.9 (Appendix B). This model includes all the covariates and had the highest adjusted  $R^2$ . For the results presented in Table B.9, the base categories of the explanatory variables are marked with an asterix (\*) in the univariate tables in Appendix B.

$$\begin{aligned}
\mathbb{E}(\text{Total score}) = & \beta_0 + \beta_1 * \text{instrument} + \beta_2 * \text{international} + \beta_3 \dots \beta_6 * \text{grade} + \\
& \beta_7 * \text{prior STAT} + \beta_8 \dots \beta_{10} * \text{class} + \beta_{11} \dots \beta_{14} * \text{gender} + \beta_{15} \dots \beta_{22} * \text{highest parent education} + \\
& \beta_{23} \dots \beta_{24} * \text{COVID engagement} + \beta_{25} \dots \beta_{26} * \text{COVID interest} + \beta_{27} \dots \beta_{28} * \text{COVID relevance} + \\
& \beta_{29} * \text{topic familiarity} + \beta_{30} * \text{topic interest} + \beta_{31} * \text{context easier} + \beta_{32} * \text{context harder}
\end{aligned}
\tag{2.1}$$

Including all covariates and survey responses to the model increased the adjusted  $R^2$  value to 25%. We also considered a model with all the variables in Equation 2.1 interacting with instrument type. Almost all the interaction terms had high p-value with very little gain in the  $R^2$  value (less than 1%), and therefore, they were not given further consideration. Results from this model are presented in Table B.10 in Appendix B.

No matter the model, instrument type was found to be related to the total score with a lower score on the modified assessment. For the model in Equation 2.1, the p-value for instrument type was 0.003. After accounting for the variation in scores explained by the instrument type, these models showed evidence of a relationship between some of the covariates and survey responses, and total score. The strongest relationships, as indicated by small p-values, were with 1) grade expectations B, C, or D (base category ‘A’; estimates -4.57, -6.88, -8.77 respectively; p-values 0.0000), 2) fourth year or higher students (base category ‘First year students’; estimate 1.77; p-value 0.08), 3) ‘Maybe’ being interested in learning about COVID-19 through a statistical lens (base category ‘Not’ interested; estimate 1.22; p-value 0.03), and 4) ‘Maybe’ considering COVID-19 to be relevant to one’s life (base category ‘Not’ relevant; estimate -3.59; p-value 0.02).

Diagnostic plots for these regression models are presented in the Appendix B. Since most of the explanatory variables we chose were categorical, scatter plots of the response variable or the residuals with the explanatory variables were not considered. The histogram of residuals looks fairly normal, with slight more density on the positive side. The scatterplot of fitted values versus residuals indicates a definitive pattern suggesting that there is an omitted variable bias in the results we are seeing. It may be reasonable to expect that additional variables at both test-taker and item level may be able to explain further variation. This may include racial and ethnic background, observed course performance, item text characteristics etc. Finally, the residual plots ascertained that heteroskedasticity is not a concern for these models.

## 2.4 Discussion

This work demonstrates two important things. First, that a carefully designed isomorphic assessment can allow for reliable measurement of statistical literacy in specific contexts. Second, that a year into the COVID-19 pandemic (as of April 2021) students who were finishing-up a semester of college-level introductory statistics scored lower on a pandemic-specific assessment of statistical literacy as compared to another version with a variety of non-pandemic contexts. This lower score indicates that context matters!

These analyses inform two distinct questions at hand and what we learn from one informs the other. On the topic of isomorphic assessments, we set out to investigate whether the M-BLIS measures the same underlying constructs as BLIS and in the same way, or not. The reliability and validity analyses indicate that this is true for most items. Though, for the items where we notice a difference in factor loading or the item

information curves, the differences are noticeable. Based on these, we can conclude that a carefully constructed isomorphic assessment can measure the same underlying constructs while exposing the test taker to statistical literacy concepts through the lens of a variety of application areas. For the items which indicate serious difference, our future work will look at factors such as reading difficulty as measured by a lexicon score that comprises of linguistic difficulty as well as length of text, whether the student is a native speaker of American english or not, whether they have prior statistics interest or not, and whether they are interested in studying the pandemic through a statistical lens or not. Responses will be analyzed to investigate which, if any, item or test-taker characteristics may be driving the differences in scores on the items with high differences in proportion of correct responses. The second question of interest to us was a comparison of student performance. These analyses were conducted assuming that scores on BLIS and M-BLIS are equatable under the internal-anchor design [193]. Various CTT-based analyses using multiple linear regression indicate that assessment type is an important predictor of total score no matter which other characteristics are included and whether the model includes any interactions or not.

### **2.4.1 Limitations**

In general, the less portable items on BLIS required either raw quantitative data, data from a randomized experiment, or data that led to visualizations with peculiar characteristics such as strong right skewness. Concessions were made in case of three items where for one item the parameter of interest was switched from mean to proportion, and an observational study was discussed instead of a randomized experiment in one other. As seen in the example in Table 2.2, reverse skewness was accepted for one item. However, the lack of open availability of raw datasets is a hurdle that will need to be addressed more systematically in creating future isomorphs.

Additionally, balancing the competing goals of maximizing engagement and minimizing emotional impact lead to the inclusion of some topics which may not be most relevant to the lives of of our target population for the study - college students, in this case - and exclusion of some topics which may be directly related to them. For example, one of the modified items referred to pre- and during pandemic performance of elementary school students on standardized tests. This issue is confounded by the expectations of the ‘college student’ audience which is typical to an educational research study, though that may not need to be the case for the general purpose of the research. The choice of the test population can bias the choice of relevant contexts.

The item in Table 2.5 was a subject of lengthy discussions, some of which included the expert reviewers. The implicit assumption of a coin being unbiased and our intuition about 50% of them landing on heads benefitted the original item. It was also based on an infinite population. However, upon deliberation, it was agreed that it is extremely hard to find other phenomena which have an unconditional 0.5 probability of occurrence which is understood intuitively, and therefore the substantial change in wording was included. The original item was an interesting case because students are assumed to be so familiar with fair coins that the frequency of their ‘encounters’ with the context might actually outweigh the other dimensions of engagement/relevance we are seeking in this study. This item may or may not be considered a true isomorph.

Authors must also acknowledge that even though we use anchor items to compare the two sets of respondents at baseline, we have to account for possible ordering effect. These identical items could function differently across BLIS and M-BLIS, especially since they may appear out-of-context on an assessment based entirely on one specific topic - the COVID-19 pandemic.

Finally, survey questions were asked at the end of the assessment. Therefore, we didn’t expect that students’ performance on the assessment would have been affected by these. However, responses to the survey questions may have contained some cognitive bias based on whether they had just seen an entire assessment based on COVID-19 or not.

## **2.4.2 Implication for Future Work**

Since the instruments are observed to function comparably, we argue that isomorphic assessment can be created to assess statistical literacy in various pertinent contexts. Even though it may be quite tedious create them, these instruments can be invaluable tools in getting respondents to consider statistics through a contextual lens that is relevant, and continue to measure how curricular strategies may affect literacy levels. Therefore, future research can be directed towards two purposes. 1) measurement of statistical literacy in various disciplinary or societal contexts using isomorphs of BLIS, and 2) using these isomorphic versions to assess performance of experimental curricular or pedagogical strategies. However, additional work exploring the transfer and cognitive processes behind statistical problem solving will also be essential to our understanding the role of contexts.

The pilot study was intended to study psychometric properties of M-BLIS in comparison with BLIS to determine whether the BLIS and M-BLIS are psychometrically

isomorphic, and whether they measure the same constructs even when the context is changed. To draw reliable conclusions, it was essential that we have the ability to compare results from our study to the field test conducted during the development of the original assessment. To achieve this, it was important to ensure that the BLIS items remained identical to that test, and therefore, M-BLIS was based on that version. At no point did we change any details in the original assessment in an effort to ensure comparability across the original work [17] and our pilot study. Resultantly, the results from this paper are specific to one definition and assessment of statistical literacy. Future research should study the role of contexts using other assessment instruments.

Differential student performance on BLIS and M-BLIS with a low p-value on inferential results indicates that the context in which a statistical question is posed affects assessment responses. Our respondents made sense of statistical questions differently based on whether the context behind the numbers was relevant to them or not. In reference to the discussion in Section 2.2.1 regarding sensitive contexts, this finding also has implications for teaching practices. If additional research finds that the sensitivity of the topic may have contributed to the lower scores on M-BLIS, an argument can be made to favor inclusion of such topics on curricular materials instead of including them in grade-affecting assessments [194].

From a context point of view, two things are worth noting. First, as discussed in Section 2.3.1, some of the BLIS items pertaining to college students saw better performance even though the examples were realistic. This may suggest that relevance itself may be hypercontextualized for different subgroups. Secondly, it was interesting to note in Table 2.11 that there was a certain percentage of students who, no matter which assessment they took, indicated after completing the assessment that they had engaged with the COVID-19 pandemic by seeking out information, believed it was relevant to their lives, yet would not be interested in gaining data-driven insights into the pandemic. Granted, this study ran about 13 months into the pandemic and there may have been pandemic fatigue. However, this was at the end of a semester during which they had taken an introductory statistics class, making this an interesting phenomena warranting further investigation.

## 2.5 Conclusion

For a statistically literate individual, the ability to marry one's understanding of statistical constructs and the context-at-hand is assumed. In fact, as there is no statistics without

context [35], statistical literacy is also inherently contextualized. However, the transfer of statistical skills to new contexts is non-trivial and this work further examines how contexts may factor into test-takers' responses to a research-based assessment of statistical literacy. Parallel to the discussion in [72] in the context of mathematics education, statistics education, too, is a way for us to develop citizens who can make sense of quantitative information in contexts that matter to them. Towards that goal, this work will allow researchers to better understand how students of statistics showcase statistical literacy skills in the context of relevant topics, and inform instructional practices which can maximize such transfer in the future.

# Chapter 3 | Effects of Teaching Through Relevant Contexts on Statistical Literacy: Evidence from a Curricular Experiment

## 3.1 Introduction

Statistical literacy as a goal of statistics education has been discussed extensively in the literature [3–5, 7–11, 13–15]. These discussions converge on the belief that a statistically literate citizen can make sense of statistical insights pertaining to topics relevant to their professional and personal life [24]. This is also critical at a societal level as highlighted by the ongoing COVID-19 pandemic [25, 26]. Previous research has also considered ways of improving statistical literacy [27–32]. The geographical diversity among researchers on this list is noteworthy. Additionally, recommendations for teaching practices which can improve statistical literacy have been discussed in guiding documents such as the GAISE college report [18, 33] and the GAISE PreK-12 Report [20]. The GAISE college report [18] recommended integrating “real data with a context and a purpose” into statistics instruction. This nudged the community to bring more real datasets into the classroom and teach through examples that may be relevant to students. There is some evidence that this helped students develop a sense for how statistics is relevant to their lives and improve their engagement with and interest in statistics [34]. The statistics education community is not unfamiliar with discussions surrounding the role of contexts in statistics and its instruction. [35] highlighted that there is no statistics without context.



Bringing these ideas together, we ask the following question in this work: Does including relevant contexts in curricular materials cause a differential gain in students' statistical literacy outcomes?

### **3.1.1 Role of Contexts in Statistics Education**

Considerable amount of work has discussed the value of contexts in statistics education and powerful ways of introducing contexts which are familiar to the students into the curriculum. [45] highlighted the centrality of contexts to statistics education as well as statistical literacy. [46] laid out the design for an entire course that focuses on real data that can develop statistical modelers and thinker. [47] posited statistical habits of mind important for learners as well as teachers, the first of them being the role of context in every stage of statistical inquiry. [49] conducted a study which found that contexts played an important role in 10<sup>th</sup> graders' development of inferential reasoning. [44] is a library of datasets and examples which facilitates the inclusion of real datasets into a variety of courses. Finally, [50] synthesized the discussions regarding and highlighted the value of guided inquiry exercises - exercises where multiple questions are built atop the same context. Research has also investigated the relationship between including contexts in curricular materials and students' engagement with and interest in statistics - as well as student outcomes.

For this work, we wanted to specifically investigate the role of relevant contexts in statistical and scientific problem solving. 5.1 conceptualize relevant contexts as ones that are societally relevant at a given time and people would have engaged with and thought about on their own. Further, they define contextualized statistical literacy as statistical literacy pertaining to relevant contexts. We use both the Basic Literacy in Statistics (BLIS) instrument developed by [17] and its modified version (M-BLIS) developed by 5.1 to measure statistical literacy as an outcome of the study. Based on the work discussed in 5.1, we treat the scores from the two assessments (BLIS and M-BLIS) to be psychometrically equivalent in our data analyses. When thinking about relevant contexts and the ways in which people's familiarity and prior engagement may play into the training and subsequent outcomes, we must also be also mindful of possible suspension of sense-making [71, 72] whereby familiarity with the context may be foregone in favor of focusing on the underlying statistical idea.

### **3.1.2 Randomized Experiments in Statistics Education Research**

As recommended by [87], randomized assignment has been discussed in statistics classrooms, starting with introductory curriculum, to highlight its importance in researchers' ability to draw causal inferences. The role of randomized experiments in educational research [88–90] and educational policy evaluations [91–93] has been discussed in the literature. However, controlled experiments can be difficult to conduct in educational settings [1, 40, chap. 3]. Randomization at the student-level can lead to interference [94] requiring randomization at the classroom or school level. Such studies can be resource-intensive due to the requirement of a large number of sections or classrooms as well as teacher training. Very few designed randomized experiments are conducted in statistics education research with a few notable exceptions. [95] randomized students to control and experimental sections to investigate the effect of simulation-based inference curricula. [96] randomly assigned separate mini readings at the student-level to examine the effects of teaching using tools that may be considered to be fun on student learning. [97] implemented a quasi-experimental design wherein two semesters of a given course enrollment were assigned treatment or control to measure the effects of teaching statistics with a critical pedagogy. [98] studied the effect of teaching through Shiny apps by assigning one of the two enrolled course sections into the treatment group.

One of the goals of the present work was to utilize research-based assessments to measure the outcomes of interest, as encouraged in [19]. Three of the four experiments discussed above implemented a similar strategy. [95] used the ARTIST (Assessment Resource Tools for Improving Statistical Thinking) topic scales [99] for specific topics of interest. [96] used two scales measuring attitudes towards statistics - SATS-36 (Survey of Attitudes Toward Statistics, [100]) and SAM (Statistics Anxiety Measure, [101]) and considered pre-test and post-test scores. [97] also gathered pre-test and post-test data on the CAOS (Comprehensive Assessment of Outcomes in Statistics, [102]) and CLES (Constructivist Learning Environment Survey, [103]) scales. Whereas, [98] used course assignments created by the instructional team.

### **3.1.3 Causal Inference in Statistics Education Research**

The three research questions stated for this project are framed as causal questions, which is an important goal of the present work. Causal conclusions can play a critical role in informing educational practices through a rigorous investigation of ideas that may or may not affect learning [89]. However, examples of research drawing causal conclusion based

on well-designed studies are scarce in statistics education literature. Of the experiments discussed above, [95] was the only study which established a causal effect of the treatment (curriculum type) on the learning outcomes. Their work analyzed the data using a multivariate analysis of covariance (MANCOVA) model. [108] used observational data to investigate the relationship between constructivist strategies in the classroom and students' attitudes towards statistics. They discussed using a causal comparative design, however, they warn against drawing causal conclusions due to the analytical strategies used. [109] also conducted an observational study to assess the effect of instructors and instructional practices on student attitudes. Outside of statistics education literature, some methodological discussions have highlighted the importance and usage of causal inference in educational studies. [110] discussed a method for estimating the causal effect of time-varying instructional treatments. [111] and [92] discussed the importance and implementation of causal-inference-based conclusions in the context of large-scale assessments in education. [112] conducted an extensive survey of various causal inference methodologies and highlight education as an important application area. [113] discussed the role of and ways to improve causal inference in educational research. The present work provides one possible framework for conducting causal analyses for statistics education research, providing a prototype for similar work in the future.

Even though an approach such as the potential outcomes framework [114,115] can be utilized to estimate causal effects, a model-based estimate is more appropriate for the present study since the treatment was not randomized at the unit-level. As discussed in Section 3.1.2, such a design choice is typical for educational studies. Data from the experiment are analyzed using a multilevel modeling strategy with covariate adjustment. The randomized experimental design employed in this work allows for causal interpretation of the coefficient of the treatment assignment. Using this approach, we address the following three research questions (RQs) in this chapter. (RQ3): does introducing relevant contexts in a statistics classroom cause a differential gain in statistical literacy outcomes? (RQ4): does taking an assessment of contextualized statistical literacy as a pre-test cause a differential gain in statistical literacy outcomes? (RQ5): does the interaction between contexts incorporated into the classroom and type of statistical literacy assessment cause a differential gain in statistical literacy outcomes?

## 3.2 Methodology

### 3.2.1 Sample

To address the stated RQ3, a randomized experiment was conducted in a co-ordinated undergraduate introductory statistics course at a large public research university in eastern United States. This course was taught under the Lock5 simulation-based inference (SBI) curriculum. This study was conducted during Fall 2021. During that semester, four faculty members taught five lecture sections. Each lecture section was divided into four or eight lab sections. Faculty members (instructors for the course) were supported by 12 graduate teaching assistants (GTAs) who conducted two (2) laboratory sessions (labs) each. Both labs assigned to a given GTA included students from the same lecture section. Every lab enrolled 80 students each, providing an initial sample size of 1960 students. GTAs were supported by one undergraduate learning assistant (LA) assigned to each section. This course offered an elegant framework for the randomization.

### 3.2.2 Tools

In this section, we discuss the several tools used to measure important variables. The intervention in this study took the form of modified lab activities and assignments (Treatment 1:  $W$ ). The content of this modification is further discussed in Section 3.2.4. A lab session was a 50-minute class period conducted following each lecture and before the next lecture. The lab included a problem-solving worksheet providing an opportunity for the students to apply ideas learned in the previous lecture to data collection, analysis, and interpretation exercises. Every student was required to sign-up for a lab section during which the GTA and the LA were available to support students' learning. A typical lab session began with a brief review of content lead by the GTA. For the remainder of the lab, students were encouraged to work in small groups or on their own, and reach out for support as needed. The lab quiz was a formative assessment including multiple choice or numerical-entry questions based on the exercises included in the lab worksheet. The lab worksheet and quiz were identical across all sections, making it the most suitable aspect of the curriculum to modify for this research. Even though it was not an important component of this research study, lab quizzes played the dual role of contributing to the final grade as well as providing frequent and low-stakes opportunities for students to understand their performance in the course.

The response variable for this study - gain score on statistical literacy assessment -

was measured using one of the two instruments discussed in Section 3.1. Each of these instruments had 37 items, each graded dichotomously for one point each. Therefore, the response variable i.e. the gain score can take whole number values between -37 and 37. Each student was randomly assigned to one of the two instruments (BLIS or M-BLIS) (Treatment 2:  $S$ ) 5.1 at the beginning of the semester and completed the same assessment as a pre-test and a post-test. In addition to the response variable, these assessments collected demographic data from the survey component as well as additional information regarding the students' interest in and engagement with relevant topics which were critical to the design of the intervention.

### **3.2.3 Experimental design**

For this experiment, half the GTAs in each lecture section were randomly assigned to treatment. The author of this dissertation was one of the GTAs and was assigned to the treated group. Treated sections completed modified lab assignments. Therefore, the treatment assignment ( $W$ ) was at lab section level. Such a design where curricular intervention is applied at classroom-level instead of student-level is common in educational studies as discussed in Section 3.1.2. It offers the advantage of avoiding interference [94] of treatment effect either through the same instructor teaching treated and untreated sections or through the mingling of treated and untreated students in the same classroom. Due to constraints pertaining to the consent procedure as required by the Institutional Review Board (IRB) protocol, student-level lab quiz scores were not available to the researchers. Only summary information (mean, standard deviation, and five-number-summary) at lab section-level was provided for this researchers. However, the primary outcome ( $Y$ ) of interest for this study was statistical literacy score measured using either BLIS or M-BLIS. Due to the random assignment of instrument type to each student, we were able to investigate two parallel causal questions. First, the effect of including relevant contexts on lab assignments on statistical literacy gain scores. Second, the effect of taking an assessment of either statistical literacy or contextualized statistical literacy at the beginning of the semester on the same gain scores.

### **3.2.4 Design of intervention**

During the semester this study was conducted, the course included 25 graded lab assignments based on course material. Each student was allowed to drop the lowest two lab scores in their final grade. Each lab worksheet comprised of 3-6 lab activities designed

to take approximately 40 minutes to solve. Due to a combination of administrative reasons, specific requests from the instructional team, or requirement for raw multivariate quantitative data, seven out of the 25 labs were modified. These seven included two out of four subsections in the chapter on confidence intervals, and all five subsections in the chapter on hypothesis testing. The modified lab activities were parallel tasks aligned as closely as possible to the original activities. It was ensured that all modified activities met identical learning goals and difficulty levels as the activities in the original curricular material, thereby ensuring that students in both groups had an equal learning opportunity, and that the treatment and control labs were comparable.

When modifying the seven labs mentioned above, two categories of activities were retained identical to the original lab activities. 1) An activity that used data collected from students in the course during the semester was unchanged. As discussed in the [18] guidelines, data about students lead to high engagement. Therefore, we did not expect any gains in engagement by replacing this with a relevant context. 2) Three activities in the hypothesis testing chapter were setup based on naked [18] hypothetical datasets and examples. The importance of the learning goals these activities met were considered to supersede the importance of including relevant contexts, especially at an experimental stage. In one case, the original lab example used blood pressure data. The modified activity retained the original example with an addition of two lines mentioning the role of blood pressure readings as indicators of health rates and the incidence of heart health concerns among young adults. Such a modification was made to three activities. Each time a relevant context was incorporated into an activity, a hyperlink directing the reader to source information was included at the end of the description or common stem.

#### **3.2.4.1 Choice of relevant topics**

As discussed in Section 3.1, an important qualifier for a context to be relevant was whether students had interacted with that topic outside of the classroom. To ensure that this was the case for the curricular modifications performed in this study, the pre-test gathered some data from respondents. A grid of survey questions asked students about their engagement with nine contexts along three dimensions. Table 3.1 summarizes responses to three statements 1) “I have actively looked for information on this topic in the last 6 months.” (Engagement), 2) “I would like to gain data-driven insights into this topic.” (Statistical interest), and 3) “I think this topic is relevant to our lives” (Relevance).

Topic	Engagement			Statistical interest			Relevance		
	Yes	Maybe	No	Yes	Maybe	No	Yes	Maybe	No
COVID-19	0.86	0.05	0.09	0.62	0.21	0.16	0.97	0.02	0.01
College student-life	0.74	0.11	0.15	0.69	0.18	0.13	0.96	0.03	0.01
Education	0.76	0.06	0.18	0.69	0.13	0.17	0.98	0.01	0
Diversity	0.47	0.17	0.36	0.61	0.22	0.16	0.92	0.05	0.03
Climate science	0.47	0.15	0.38	0.62	0.19	0.19	0.85	0.11	0.04
Immigration	0.33	0.16	0.51	0.52	0.25	0.24	0.78	0.15	0.06
Mental & physical health	0.75	0.07	0.19	0.79	0.11	0.1	0.98	0	0.02
Politics & governance	0.68	0.13	0.19	0.58	0.16	0.26	0.89	0.07	0.04
Healthcare advances	0.54	0.12	0.35	0.64	0.18	0.18	0.92	0.06	0.03

Table 3.1: Survey questions regarding various contexts

Based on this information, we focused on the topics with the highest proportion of engagement (COVID-19 pandemic, Education, Mental and Physical Health, and College-student life). Contexts at the intersection of multiple topics, such as COVID-19 vaccination rates among college students, were. The remaining five topics were also included where applicable. However, they appeared less frequently than the four topics mentioned above. Some examples of modified contexts included in the experimental labs include,

- a bill in congress about making college tuition-free in the US,
- diversity index and vaccination rates for states in the US,
- energy production generated from renewable sources,
- academic distress and status as a first-generation college-student,
- approval for interracial marriages, and

Another important consideration in determining whether the context was relevant or not was the ‘current’-ness. Although all the contexts described above and considered in the study can be considered relevant at a societal level, in order to emphasize this to the students, attempt was made to find data collected within no more than 3 years prior to the study and the year of data collection was recorded explicitly in the document provided to the students.

### 3.2.4.2 Example

An example of a lab activity is demonstrated in Table 3.2. This was the last lab in the hypothesis testing chapter and was designed to help students develop and practice their understanding of type I and type II errors.

Original item stem	Modified item stem
We are testing a new drug with potentially dangerous side effects to see if it is significantly better than the drug currently in use. If it is found to be more effective, it will be prescribed to millions of people.	In April 2021, FDA and CDC recommended a pause on the Johnson & Johnson COVID-19 Vaccine in the US due to a potentially dangerous side effect. Follow-up analysis was conducted to determine if it is significantly better to administer the vaccine than not. If the vaccine was found to be safe and effective in preventing COVID-19, it would be administered to millions of people. (Source: CDC)
Now we are testing to see whether taking a vitamin supplement each day has significant health benefits. There are no (known) harmful side effects of the supplement.	The National Institutes of Health in the US considered evidence to determine whether taking a vitamin C supplement each day significantly reduced the time to recovery in COVID-19 patients who are NOT severely ill. There are no (known) harmful side effects of the supplement. (Source: NIH)

**Table 3.2.** Example lab activity

For each of the two scenarios, students were required to interpret both errors in the context of the problem and comment on which error would be considered to be worse than the other.

### 3.2.5 Causal inference methodology

This experiment was conducted with the intention of inferring the causal relationship between inclusion of relevant contexts in the curriculum (treatment or intervention  $1$ ,  $W$ ) and statistical literacy change score from beginning to the end of the semester (outcome  $Y$ ) (RQ3). Educational literature contains some discussions regarding using gain scores as an outcome versus using post-test scores with pre-test scores adjusted for in the model. Gain scores have been discussed as more reliable [195–197]. Gain scores were also substantively more interesting. Even though errors on pre-test and post-test may be correlated, potentially violating model assumptions, gain scores were chosen as the outcome for this study. For the present work, the causal effect is the



difference between the statistical literacy change score that would have been observed if a student was assigned to complete lab exercises based on relevant contexts and the change score for the same score if the student was assigned to the control group. Due to the hierarchical nature of this experiment, we pursued the multilevel modeling approach to causal inference which can allow for interpretation of the causal effect under the potential outcomes framework [198, 199]. We started with a simple model (Equation 3.1) which estimated the parameter for the effect of the treatment ( $W$ , at section-level) and specified random effects for the four instructors and the 24 lab sections to account for the variability within them. This model matched the nested data structure most closely.

$$\begin{aligned}
\forall i = 1, 2, \dots, n \text{ (students)}, j = 1, 2, \dots, m \text{ (lab sections)}, \text{ and } k = 1, 2, \dots, l \text{ (instructors)}, \\
y_{kji} &\sim \mathcal{N}(\alpha_{kj[i]}, \sigma_y^2) \\
\alpha_{kj} &\sim \mathcal{N}(\alpha_k + W_j\theta, \sigma_{\alpha_k}^2) \\
\alpha_k &\sim \mathcal{N}(\alpha_0, \sigma_\alpha^2),
\end{aligned} \tag{3.1}$$

where,

$y_{kji}$  is the outcome for student  $i$  taught by instructor  $k$  and enrolled in lab section  $j$ ,  $\alpha_{kj}$  is the true mean gain score for section  $j$  from instructor  $k$ 's lecture section across all students,  $\theta$  is the causal effect at section level based on treatment assignment  $W$ , and  $\alpha_k$  captures the random effects at the instructor level (need to modify model to capture section-level random effect).

Next, person-level pre-treatment covariates were added to the mixed effects or hierarchical linear model (HLM) to estimate the causal effect in the presence of important covariates. No lab section-level or instructor-level covariates could be added on account of design choices which led to perfect collinearity across covariates and the random effects. Visual summaries were considered before inferential results to assess covariate balance across the treatment and the control group. Checking this balance informed our understanding of whether any of the covariates were essential to adjusting for pre-treatment differences across the two groups. Equation 3.2 shows the full model with varying-intercepts for instructors and lab sections in addition to fixed effects for person-level covariates, in addition to the treatment effect coefficient.

$$\begin{aligned}
&\forall i = 1, 2, \dots, n \text{ (students)}, j = 1, 2, \dots, m \text{ (lab sections)}, \text{ and } k = 1, 2, \dots, l \text{ (instructors)}, \\
&y_{kji} \sim \mathcal{N}(\alpha_{kj[i]} + \mathbf{X}_{kj[i]}\beta, \sigma_y^2) \\
&\alpha_{kj} \sim \mathcal{N}(\alpha_k + W_j\theta, \sigma_{\alpha_k}^2) \\
&\alpha_k \sim \mathcal{N}(\alpha_0, \sigma_\alpha^2),
\end{aligned} \tag{3.2}$$

where,

$y_{kji}$  is the outcome for student  $i$  taught by instructor  $k$  and enrolled in lab section  $j$ ,  $\beta$  are coefficients on student-level covariates such as statistical literacy pre-test score, and demographics and survey questions collected on the pre-test,

$\alpha_{kj}$  is the true mean gain score for section  $j$  from instructor  $k$ 's lecture section across all students,

$\theta$  is the causal effect at section level based on treatment assignment  $W$ , and

$\alpha_k$  captures the random effects at the instructor level (need to modify model to capture section-level random effect).

Even though it was not the primary goal of this study, the randomization of BLIS and M-BLIS across all study participants could be treated as a secondary treatment variable. It is conceivable that students completing an assessment of statistical literacy based on relevant contexts (M-BLIS) may approach statistics differently during the course learning period in comparison to those who complete the same assessment with other mixed contexts at the beginning of the course. Therefore, we fit models in Equation 3.1 and Equation 3.2 separately for the BLIS and the M-BLIS group, and then modified them to replace  $W$  with assessment type ( $S$ ) as the treatment variable (RQ4). Additionally, we also considered an interaction between the two treatment variables ( $W$  and  $S$ ) for each of the two models (RQ5).

### 3.3 Results

In this section, we discuss findings from the models presented in Section 3.2.5. Corresponding descriptive analyses are also presented alongside the inferential results. Even though the course enrollment was 2000 students, the final analytical sample size was  $n = 265$  due to attrition at several levels. Each student was randomly assigned to complete either BLIS or M-BLIS for both the pre-test and the post-test. The number of

students who consented to participation and attempted the assessment were as follows: 1,015 completed pre-test (496 BLIS responses and 519 MBLIS responses) and 1,085 completed post-test (546 BLIS responses and 539 MBLIS responses). For the purpose of this study, only full assessment responses were considered. Removing incomplete responses resulted in 715 responses on the pre-test (357 BLIS, 358 MBLIS) and 937 responses on the post-test (483 BLIS, 454 MBLIS). However, only 305 of these responses could be matched across pre- and post-test based on a unique six-digit alphanumeric ID since assessment responses were anonymized. Finally, students were required to select their lab section in the survey component of the assessment. Since the treatment was applied at lab-section level, it was considered crucial to ensure that the matched pre-post test responses were in the same section. 40 students did not meet that criterion, resulting in a sample size of 265.

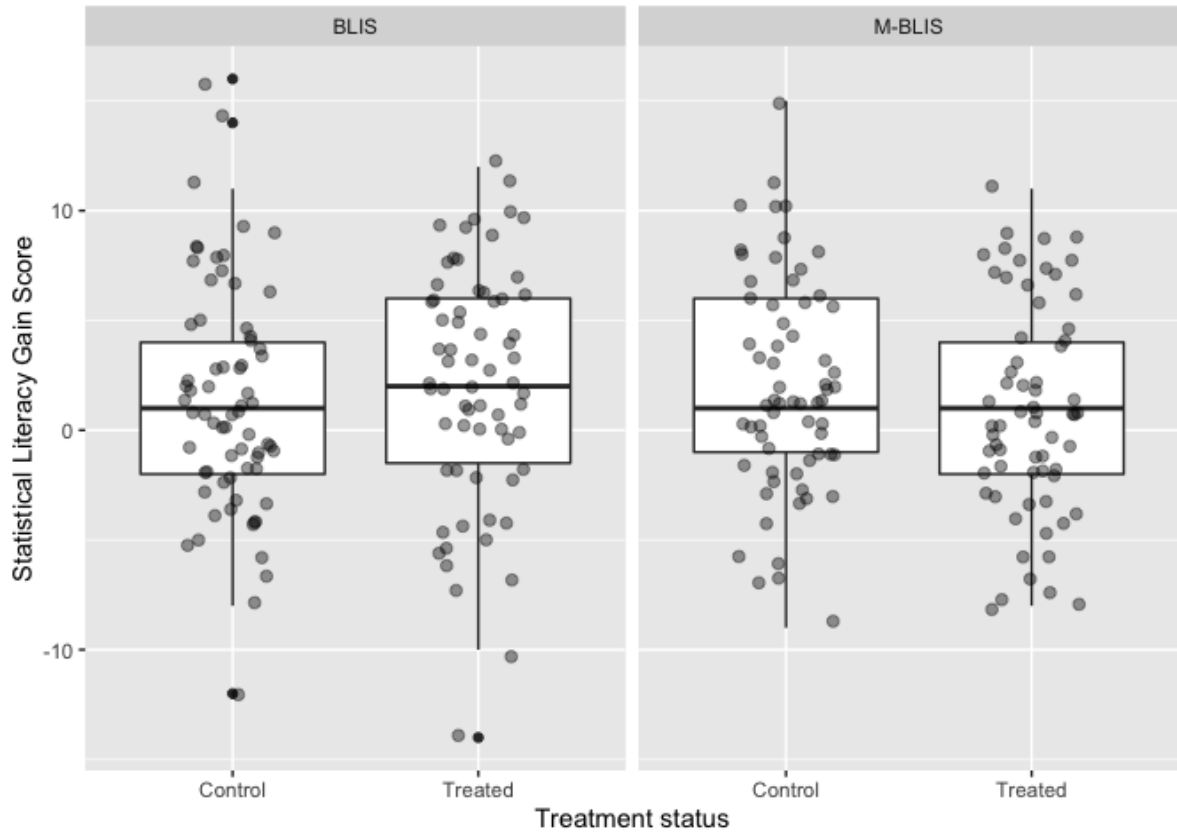
### 3.3.1 Mixed Effects Models

As discussed in Section 3.2.5, the first goal was to consider the causal effect for the two possible treatment variables - lab modification at lab section-level (RQ3) and type of assessment at individual student-level (RQ4), as well as the their interaction (RQ5). Table 3.3 and Figure 3.1 summarize the distribution of the response variable (gain score on statistical literacy) vis-a-vis the treatment variables.

	Mean	Median	SD	IQR
Control	1.63	1	5.01	7
Treated	1.50	1	5.14	8
BLIS	1.70	2	5.26	7.50
M-BLIS	1.42	1	4.87	7.75

Table 3.3: Summaries of gain scores

Mean and median gain scores across the treated and control groups were very close to each other. Their spreads were also comparable, with slightly higher scores in the control group. When looking at the type of assessment as a treatment variable, mean and median scores were higher on BLIS than M-BLIS. However, further analyses accounting for the varying spreads provided additional information regarding the possible difference since the range was approximately 6 points lower for gain scores on M-BLIS. As per 5.1, the overall scores on M-BLIS were lower than BLIS. Figure 3.1 considers the interaction



**Figure 3.1.** Boxplot of gain score for both treatments

between the two treatment variables. The medians were similar across all four subgroups, with slight variations in spreads. Sample sizes were comparable across subgroups - 65 (Control) and 69 (Treatment) in the BLIS group and 65 in both M-BLIS groups.

For this study, we stipulated that the variation at instructor as well as lab-section level would be important to account for. However, since they could be considered to be a random sample of instructors and lab groupings, we were not interested in their fixed effects. Figure C.1 (Appendix C) shows distribution of gain scores by instructor and Figure C.2 (Appendix C) by lab section. Both plots suggest that there would be benefit to including random effects for both these pre-treatment covariates. Having said that, the small sample sizes within lab sections (noted within parantheses under the section number) are important to keep in mind.

### 3.3.1.1 Causal effects without covariates

Table 3.4 displays estimates from the model specified in Equation 3.1. Treatment effects were estimated from three separate models with the following treatment variables: 1) modified lab section ( $W$ ) - RQs3, 4) assessment type ( $S$ ) - RQ3, and 5) an interaction between  $W$  and  $S$  - RQ5.

	Model 1 - $W$		Model 2 - $S$		Model 3 - $W*S$	
	Estimate	t-value	Estimate	t-value	Estimate	t-value
Intercept	1.64	2.81	1.66	2.88	1.33	1.90
Treatment effect - Treated	-0.20	-0.31			0.67	0.73
Treatment effect - M-BLIS			-0.25	-0.40	0.63	0.72
Treatment effect - Treated*M-BLIS					-1.75	-1.40

Table 3.4: Causal effects

Due to the difficulty in determining degrees of freedom for hierarchical linear models, t-values are used to comment on the strength of evidence against the null hypothesis of no treatment effect. Using the cutoff value of 2 for the t-value, we do not have sufficient evidence to claim that individual treatment effects were different from zero for either RQ3 or RQ4. Though lower than 2, the interaction term in Model 3 had a t-value of  $-1.4$  suggesting further investigation of the relationship between the interaction of the two treatments and gain scores. Though this evidence was weak, it was stronger than Models 1 and 2. The residual variance after accounting for random effects was very close across the three models.

### 3.3.1.2 Causal effects with covariates

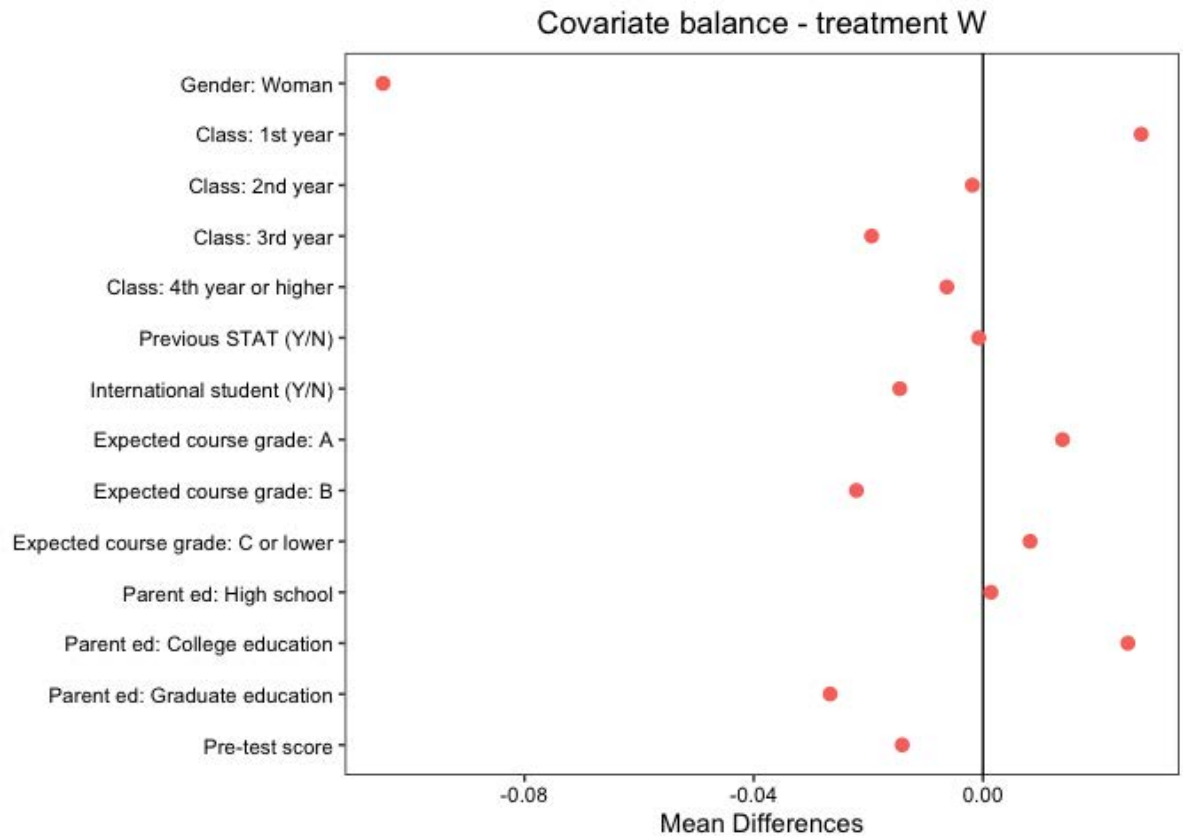
As discussed in Section 3.2.5, the model without covariates was extended to include various individual-level pre-treatment covariates. Figures C.3 - C.9 (Appendix C) capture the relationships between gain scores and these variables. Other than the indicator for whether a student has previous statistics training, all other categorical variables showed some differences in gain scores across categories, either in their medians or spreads. The scatterplot of pre-test scores and gain scores (Figure C.9 in Appendix C) indicated higher variance in gain score for higher pre-test scores.

Five options were provided for the question ‘What gender do you identify as?’ (Figure C.3 in Appendix C) They were Woman (1), Man (2), Transgender (3), Prefer to self-

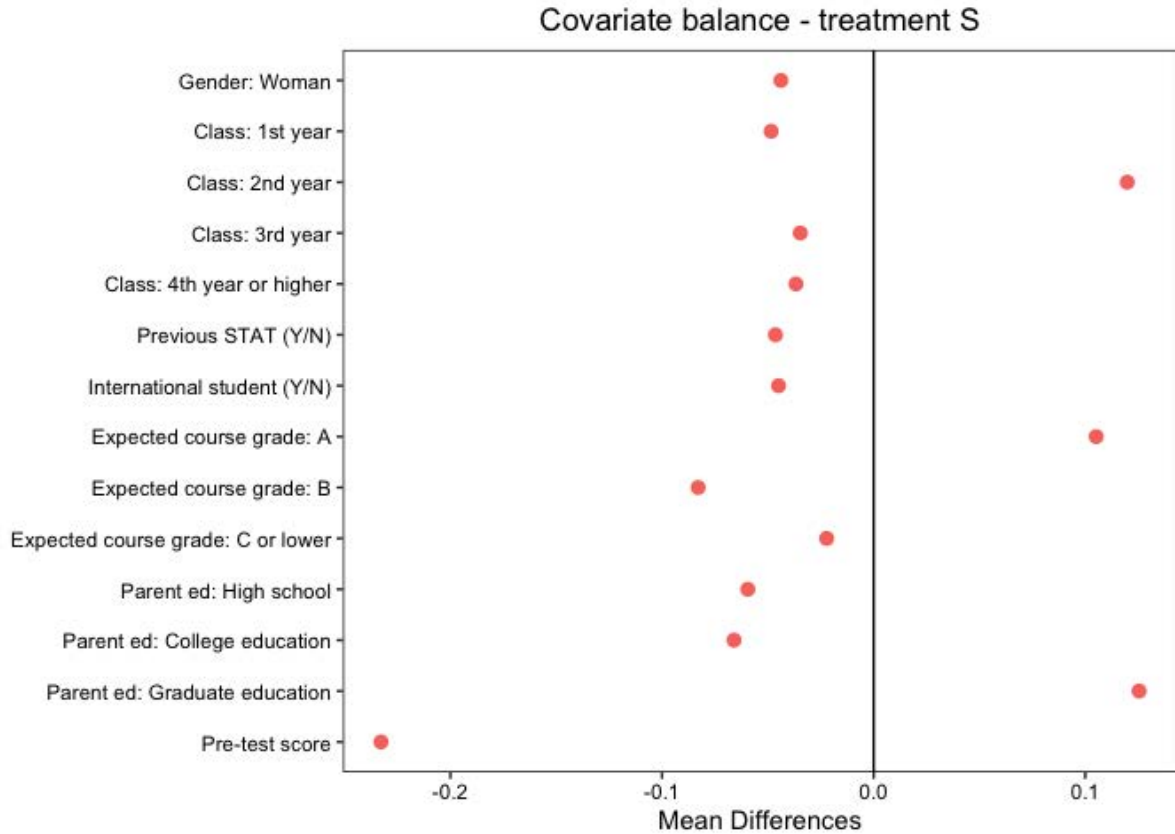
specify (4), and Prefer not to disclose (5). In the the final analytical sample, nobody selected option (3), and options (4) and (5) were selected by one person each. Due to the single observations, those two observations were removed from the data used for the analyses presented in Section 3.3.1.2.

Information about “the highest level of education received by a parent/guardian in the household” was collected at a more granular level than what is presented in Figure C.8 (Appendix C). However, 9 total categories of educational level were collapsed into 3 for ease of interpretation. Each of the final three categories includes either incomplete or complete education at the given level. For example, ‘College education’ includes ‘Some college, no degree’, an ‘Associate’s’, or a ‘Bachelor’s’.

Love plots were plotted for the two treatment variables (Figures 3.2 and 3.3) to look at covariate balance across the treatment and the control group. This is an important step in causal analyses and allows us to determine a-priori whether including a variable in the model is essential to adjusting for pre-treatment differences across the two groups or not. Pre-test scores was the only quantitative variable, and therefore, the mean difference is standardized for that variable. For all other variables, raw differences are displayed. When considering the main experimental treatment (modified lab sections -  $W$ ), categories of the gender variable had most extreme mean differences. Though at the middle-school level, [53] had found gender-based differences in statistical literacy on a separate instrument. Whereas, for the instrument type as a treatment variable ( $S$ ), pre-test score had the highest absolute mean difference along with specific levels of certain categorical variables showing mean differences slightly above 0.1. It is worth noting, though, that for the scale of the response variable, these differences are relatively small.



**Figure 3.2.** Love plot for modified lab as treatment



**Figure 3.3.** Love plot for instrument type as treatment

Table 3.5 shows estimates from the model specified in Equation 3.2. Once again, we fit three separate models, each of them including a distinct treatment effect specification. However, the covariates as well as the random effect specifications (instructors and lab-sections) were identical for all models.

	Model 1 - W		Model 2 - S		Model 3 - W*S	
	Estimate	t-value	Estimate	t-value	Estimate	t-value
Intercept	4.55	2.37	4.98	2.60	4.56	2.26
Treatment effect - Treated	-0.16	-0.25			0.67	0.71
Treatment effect - M-BLIS			-0.58	-0.90	0.31	0.35
Treatment effect - Treated*M-BLIS					-1.67	-1.28
Gender: Woman	0.31	0.41	0.24	0.33	0.29	0.39
Class: 2nd Year	0.44	0.52	0.53	0.63	0.58	0.67
Class: 3rd Year	0.92	0.81	0.98	0.88	0.90	0.80
Class: 4th or higher	0.72	0.48	0.66	0.44	0.71	0.47



	Model 1 - W		Model 2 - S		Model 3 - W*S	
	Estimate	t-value	Estimate	t-value	Estimate	t-value
Previous STAT: Yes	-0.10	-0.14	-0.03	-0.04	-0.01	-0.02
International student: Yes	0.49	0.26	0.17	0.09	0.34	0.18
Expected course grade: B	-1.02	-1.48	-1.04	-1.52	-0.93	-1.33
Expected course grade: C or lower	-3.09	-1.52	-2.94	-1.46	-3.21	-1.57
Parent ed: High school	1.50	1.07			1.09	0.77
Parent ed: Graduate education	0.06	0.10			0.11	0.16
Pre-test score	-0.18	-2.25	-0.19	-2.33	-0.20	-2.42

Table 3.5: Causal effects - with covariates

After accounting for covariates, none of the treatments showed strong evidence against the null of no treatment effect. Similar to previous analyses, we used a t-value of 2 as a reference. The pre-test score had a high t-value with negative estimates across all the models suggesting that higher the pre-test score, lower the gains. Given the nature of the relationship between pre-test scores and gain scores, this seems appropriate. When looking at students' self-reported expected course grades at the beginning of the semester, those who were expecting a grader lower than A (B, C, or D) showed some evidence of lower gains than those expecting an A. With a baseline at those students whose parent or guardian with the highest education had college education, those who's parents had only high school education showed weak evidence of higher gains. However, due to a singularity in fitting, this covariate was dropped from the second model.

Finally, an interaction model which included all variables specified in Model 3.2 with the addition of an interaction with the treatment status was considered. Parameters were estimated for such a model with both treatment variables -  $W$  and  $S$ . The RQ of interest was: does the treatment cause a differential change in gain score across subgroups? All groups with small sample sizes ( $n \leq 10$ ) were excluded from this comparison. We focused on treatment effects which reversed in direction across the treatment and the control group and had a t-value in the vicinity of or higher than 1.5 for the reversal to show some evidence of differential effect. For the lab modification treatment ( $W$ ), none of the effects satisfied these criteria. However, when treating instrument type ( $S$ ) as a treatment variable, two effects were worth noting. 1) Over and above those who identified as men, women's gain scores were lower on BLIS (estimate -0.86 with t-value -0.77) than on M-BLIS (estimate 2.27 with t-value 1.5). 2) Over and above those who had college

education as highest level of parental education, students with a parent/guardian who went to graduate school had a higher gain score on BLIS (estimate 1.25 with t-value 1.30) than on M-BLIS (estimate -2.58 with t-value -1.89).

## **3.4 Discussion**

This investigation of whether the inclusion of relevant contexts into teaching materials affects statistical literacy levels makes two key contributions despite the inconclusive results. First, the design of the study allows for rigorous causal analysis in a statistics education research study. Second, using research-based assessments to measure outcomes facilitates comparison with other similar studies which may be conducted in the future.

The broad benefits of this study can be long-term and impact statistics education practices more broadly. It is an implicit assumption that students develop an ability to apply learnings from the course to important contexts encountered in their lives outside the classroom as well as contexts impacting the world around them. This assumes knowledge transfer [65–67] and warrants further study to investigate the nature and the extent of this transfer. Further, as discussed in Section 3.1.1, there has been great push in the statistics education community to bring real data into the classrooms. However, real data need not always be relevant. Connecting contextualized statistical literacy to the types of data included in the course materials will provide us insight into whether this ability can be further improved, informing future statistics educators.

### **3.4.1 Limitations**

There was three ways in which the design of the intervention may have limited the effects observed in the study. First, this study was conducted during first fully in-person semester following the COVID-19 pandemic. Resultantly, students were not required to attend the lab sessions during which they could receive support from the TA and LA, as well interact with students within their section. This could have led to interference due to students working with fellow classmates from other lab sections. Second, since only a subset of the lab activities was modified for the treatment, that may have limited the effect size. Finally, it is worth noting that the modification of lab activities to include relevant contexts was only a small part of the curriculum. All other components of the course including the unmodified labs, lecture materials, common homework assignments, and exams were unchanged. Over the course of past several years, the instructional

team for this course had undertaken concerted efforts to make this course relatable and relevant to students. This could have potentially limited the size of the treatment effect.

On the outcome side, [17] discussed the validity of BLIS as a mid-term and end-term assessment. M-BLIS was only deployed as a post-test in 5.1. Therefore, the use of these assessments as a pre-test in the present study should be treated with care. However, the design of BLIS is such that none of the items are invalid as a pre-test. Relatedly, gain scores have two key limitations which should be considered. First, gain scores lead to boundary effects for respondents who perform either very well or very poorly on the pre-test. The low gain scores of respondents with high pre-test scores are due to the upper bound and may get interpreted as low effects of the treatment. Figure C.9 displays the gain scores on the Y-axis and pre-test scores on the X-axis for the present study. The weak negative correlation ( $r = -0.13$ ) captures the overall direction of this limitation. Second, using gain scores as the response variable requires that only those respondents who responded to both the pre-test and the post-test can be used in final analyses. This may lead to loss of information. However, for the present study, the correlation between pre-test and post-test scores for matched responses was  $r = 0.55$ , also captured in Figure C.10. In such a situation where pre-test and post-test scores are moderately correlated, the matched samples may offer an opportunity for reliable inference of causal effects, even though the sample may no longer be representative.

Finally, the final dataset in this study is approximately 10% of the enrollment in the course. Even though Missingness At Random is assumed for the attrition, that is unlikely to be the case. Section 3.3 outlines the stages of attrition. Therefore, it is difficult to estimate whether larger retention could have lead to different results despite the comparable group sizes in the current dataset. We also acknowledge the limited generalizability of these results since the experiment was conducted at a single institution and in a single course. However, this is typical of educational studies and future work will be aimed at multi-site studies.

### **3.4.2 Implications for research**

Future researchers interested in investigating the impact of teaching through contexts relevant to students must consider several important aspects. First, the list of topics this intervention worked with was pre-specified by the researcher. It would be worthwhile to build this list using open-ended responses from students themselves to include topics students report as engaging and relevant. Second, student engagement with topics can differ based on the component of the course in which they are incorporated. Future work

should consider the possible differential effect of modifying other parts of the curriculum and assessments. Finally, contextualized statistical literacy is one possible measure of someone's ability to make sense of statistics pertaining to relevant contexts. However, it is possible to conceive of other ways of assessing the transfer of classroom learning to topics relevant to individuals. With the larger goal of promoting a statistically literate citizenry, other methods of measuring the abilities deserve merit as well.

On the methodological side, this study establishes the feasibility of and a framework for conducting studies which can allow for causal claims to be made in educational research. Such a pursuit is not without its challenges such as infeasibility of randomized assignment at an individual-level, buy-in from the administration, consent procedures required by individual institutional review boards, and ensuring equal learning opportunity to the experimental and control group. Many of these challenges can be addressed by conducting such a study in an online setting. However, more research-based causal claims regarding the effects of instructional ideas on important learning outcomes are critical to making evidence-based teaching decisions. Therefore, future statistics education research should consider conducting more randomized experiments or carefully designed observational studies and drawing careful causal conclusions based on those.

### **3.4.3 Implications for teaching**

5.1 comment on the negotiation vis-a-vis sensitive relevant contexts and the possible advantage of including them on curricular material as opposed to assessment items. Inclusion of such context in teaching materials can have an added advantage that no matter the statistical literacy outcomes, the course materials encourage students to consider topics relevant to the society from a statistical point of view. However, additional considerations such as attitudes towards statistics and engagement with the subject are also equally important. If future studies lead to findings consistent with this work, contexts in a statistics classrooms should be chosen to optimize outcomes other than statistical literacy.

## **3.5 Conclusion**

[200] unequivocally asserted the mantra that not only should we assess what we value, but we must also teach what we wish to assess. We need to first outline the learning outcomes we value, create an assessment plan, and then, consequently, design curriculum

to teach what we intend to assess. As discussed in Chapter 1, statistics educators value students' ability to become statistically literate citizens who can make sense of data-driven information pertinent to their personal and professional lives. In Chapter 5.1, we discussed a proposal to assess such an ability - contextualized statistical literacy - using a research-based tool. The present work contributes to literature on the final step of this cycle, designing teaching materials intended to improve statistical literacy and evaluate its effects on the intended outcome. This work will allow researchers to further consider the effects of incorporating relevant contexts into curricular materials. More importantly, though, it will encourage more statistics education research that evaluates the effects of curricular and pedagogical decisions on learning outcomes through rigorously designed randomized experiments and appropriate causal inference methodology.

# Chapter 4 |

## Application of Cognitive Diagnostic Modeling to Statistical Literacy

As discussed in Chapter 1, a cognitive diagnostic model assesses test takers' ability vis-a-vis latent cognitive skills (LCSs) which are required to answer test questions. The inputs into this model are 1), a dichotomous item-response matrix based on observed responses ( $X_{ij}; i = 1, 2, \dots, I$  persons, and  $j = 1, 2, \dots, J$  items) and a binary Q-matrix [119] ( $Q_{jk}; j = 1, 2, \dots, J; k = 1, 2, \dots, K$ ) specifying whether skill  $k$  is needed to answer item  $j$ . The model has two key components. The first component is the Item Response Function (IRF) specifying the probability for person  $i$  to answer item  $j$  correctly depending on the skills needs for the item and possessed by the person. The second component is the Joint Attribute Distribution (JAD) specifying the joint distribution of all skills specified in the Q-matrix. Through a variety of parameters estimated from this model, the key focus is three quantities - 1) skill distribution i.e. the proportion of respondents with a given skill, 2) skill class distribution i.e. the proportion of respondents possessing a specific combination of all skills, and 3) individual skill profiles i.e. which skills does a given respondent possess.

The aim of a cognitive diagnostic model is to ensure that the test can provide diagnostic feedback on their strengths and weaknesses on these skills. In general, it can be beneficial to think of LCSs as attributes [120] because in their broad capacity, CDMs are also utilized for psychological health assessments. [121] mention that an "attribute may include procedures, heuristics, strategies, skills, and other knowledge components." In those cases, a diagnosis of whether an individual possesses a certain attribute or not can be useful for diagnostic purposes. However, for the purpose of this work which

focuses on an assessment of statistical literacy, we will continue to use the term skill. As mentioned earlier, the inputs into this model are 1), a dichotomous item-response matrix based on observed responses ( $X_{ij}; i = 1, 2, \dots, I$  persons, and  $j = 1, 2, \dots, J$  items) and a binary Q-matrix [119] ( $Q_{jk}; j = 1, 2, \dots, J; k = 1, 2, \dots, K$ ) specifying whether skill  $k$  is needed to answer item  $j$ .

In this chapter, we analyze data from the pilot study conducted in using cognitive diagnostic models to address the following research questions. (RQ6): Over and above the component skills identified as important for answering questions on the assessment for statistical literacy, does a latent ‘context familiarity’ affect the probability of correctly answering the items? (RQ7): Can the modified assessment of contextualized statistical literacy (MBLIS) provide feedback on the same statistical skills as BLIS?

## 4.1 Methodology

### 4.1.1 Q-matrix

[126] discuss the limited availability of expert-specified Q-matrices and [141] pose this requirement as a barrier to more widespread applications of cognitive diagnostic modeling. While we acknowledge this limitation, this work took the opportunity to develop a Q-matrix for measuring statistical literacy skills using BLIS and MBLIS was formed by the research team. The full matrix is available in Appendix D. The final matrix included seven statistical skills and one context-familiarity skill, as listed below.

1. *CommunicateInterpret*: Communicate/interpret statistical results.
2. *Descriptive*: Answer a statistical question based on descriptive statistics.
3. *Inferential*: Answer a statistical question based on inferential statistics.
4. *Visualizations*: Answer a statistical question based on visualizations.
5. *Univariate*: Answer a question based on univariate statistics/information.
6. *Bivariate*: Answer a question based on bivariate statistics/information.
7. *StudyDesign*: Understand study design in order to answer a statistical question.
8. *ContextCOVID*: Be familiar with the context - COVID-19 - an item pertains to.

As a starting point to develop this matrix, we considered the definition of statistical literacy which BLIS is based on - ‘Statistical literacy is the ability to read, understand, and communicate statistical information.’ It was stipulated that every item on the instrument would require the respondent to read and understand statistical information to answer it correctly. An attempt was made to codify ‘read and understand’ in form of a cognitive skill by considering textual length of item stem, whether numerical information is presented in the stem or not, and whether calculation needs to be conducted in order to consider answer choices. However, this skill was not included in the final Q-matrix due to the lack of formal criteria. ‘Communication’ was interpreted as interpretation in the context of this assessment. All items which required the respondent to choose the ‘correct’ interpretation statement from among answer choices was considered to require this skill.

In the second step of developing the Q-matrix, topics and learning outcomes detailed by the original assessment developer (Table A.1) were considered since they capture the abilities of a statistically literate citizen as per this assessment instrument. Several key features of items were noted based on this original blueprint. Any learning outcome with the word ‘interpret’ in the description was considered in conjunction with the ‘communication’ criteria based on the definition discussed above to determine whether the *CommunicateInterpret* skill is needed for the given item.

Relatedly, an important consideration in all interpretation tasks was whether the respondent needed to interpret a descriptive quantity or results from an inferential procedure. This criteria was broadened to capture skills 2 and 3 listed above - *Descriptive* and *Inferential*. The ‘descriptive or inferential statistics’ broadly include ‘descriptive or inferential procedures’ to clarify that a numerical or other output may not be presented in the item. For example, an item that required an interpretation rejecting the null hypothesis or the hypotheses themselves was considered to require the *Inferential* skill. Items regarding samples and populations, variable types, and type of study design were deemed to not require the mastery of either of these two skills so long as a statistical procedure was not central to the question. On the other end of the spectrum, items which required the respondent to consider an inferential procedure in tandem with a descriptive statistics were considered to require both these skills.

Some items included visual information and the ability to work with such information was considered an important skill. Even though determining whether or not each item needed mastery of this skill was relatively easy, some items featured visual information that is not a typical statistical visualization. For example, an item described confidence



intervals as simple line plots and required the respondent to consider how the length of this line will be affected by the confidence level. This item was considered to require the said skill.

In the final stage of developing the Q-matrix, we took a close look at the modified test blueprint (Table A.2) developed for MBLIS as well as the items themselves to consider any additional skills may be required to answer it correctly. This led to the addition of the *Univariate*, *Bivariate*, and *StudyDesign* skills. In the first draft of the Q-matrix, the skill to answer a statistical question based on multivariate information was also considered. However, only one item stem on MBLIS referred to a survey which collected information on multiple variables, and a respondent was not required to consider any of the variables themselves. Therefore, this skill was not retained in future drafts. Determination of whether an item works with univariate or bivariate information was easy.

Finally, the *StudyDesign* skill considered whether a respondent needed to consider the study design decisions in order to fully understand the question and answer it correctly. Study design decisions include, but are not limited to, whether a random sample was collected, whether random assignment was made, or whether repeated measures were collected. Care was needed in determining whether aspects of study design mentioned in item stem were relevant to the task and response choices themselves. Mastery of this skill was only considered necessary for items in which design choices were directly tied to answering the question.

Finally, the *ContextCOVID* skill was central to the investigation in this study. MBLIS measures contextualized statistical literacy . Therefore, an underlying assumption is that familiarity with the context to which an item refers is important in answering an item correctly and thereby the assessment of whether a respondent is statistically literate in that context. Since MBLIS was developed with the COVID-19 pandemic as the central context, all isomorphic items were considered to require this skills. The six anchor items and a seventh item which essentially remained unchanged are marked zero in the context skill column.

### **4.1.2 Analytical approach**

The MBLIS response data were analyzed under the CDM framework to investigate the role of a context familiarity skill in responding to an assessment of contextualized statistical literacy. The Q-matrix discussed in Section 4.1.1 proposed a measurement framework for statistical literacy. Specifically, the first seven skills captured the statistical aspects of statistical literacy and we wanted to understand how incorporating the context

skill over and above those seven affected model estimates. In the first set of analyses we inquired whether any difference is observed in the probability of answering questions correctly if context familiarity is assumed to be required for those items. In the second set of analyses, we focused on the estimates of skill prevalence, specifically the context skill, to determine whether the measurement on context-specific assessment (MBLIS) changes the estimates of skill prevalence. All analyses are conducting using the *GDINA* package [127] in *R*.

#### 4.1.2.1 Context Skill and Item Response Probabilities

We fit two DINA models (Section 1.2.2.2.4) to the MBLIS response data - one based on a Q-matrix that included the context skill and another without that skill. DINA is a non-compensatory model where the probability of answering an item correctly depends on mastering all required skills as specified in the Q-matrix. The Item Response Function (IRF) for this model is as follows:

$$P(X_{ij} = 1|a^i = \alpha^c) = \delta_{j0} + \delta_{j12\dots K} \prod_{k=1}^K (Q_{jk} \times \alpha^c_k), \quad (4.1)$$

where  $K = 8$  for the model which includes the context skill and  $K = 7$  for the other model. Since the main research question of interest (RQ6) pertains to item response probabilities, we focus on the  $\delta_{j12\dots K}$  coefficient of the full interaction, and refer to it as the AND  $\delta$ . This coefficient is the change in probability of answering item  $j$  correctly over and above the intercept for an individual who has all the required skills. This probability is  $1 - s_j$  where  $s_j$  is the probability of slippage specified in Equation 1.8. Therefore, the AND  $\delta$  is the probability that someone with all the requisite skills answers the item correctly.

Our conceptual stipulation was that the context skill is needed in conjunction with other statistical skills in order to answer a context-specific question correctly. The non-compensatory DINA model accounts for this constraint. However, DINA does not consider the effect of individual skills and even if an individual lacks one of the multiple required skills, the individual is not considered to be equipped to answer the item correctly. Therefore, we fit the A-CDM model (Section 1.2.2.2.6) which includes only the main effects for each required skill in the IRF stated as

$$P(X_{ij} = 1 | a^i = \alpha^c) = \delta_{j0} + \sum_{k=1}^K \delta_{jk} (Q_{jk} \times \alpha^c_k). \quad (4.2)$$

This model was only fit to the data with the full Q-matrix - including the context skill - to test for whether  $\delta_{j8}$  for the context skill was significantly different from zero  $\forall j$ . Since the existing modeling framework does not allow for setting these coefficients to be equal, we needed to conduct simultaneous tests for 30 coefficients, one for each eligible item. Multiple testing was conducted with bonferroni correction using the standard errors provided by the package, assuming normality. For significance level  $\alpha = 5\%$ , the bonferroni adjusted  $\alpha_{corrected} = 0.05/30 = 0.0017$ . The p-value in this case was the probability that a randomly observed  $\delta_{j8}$  estimate would be as extreme as or more extreme than the model estimate, given the standard error from the model, assuming that  $\delta_{j8}$ s follow a normal distribution. The p-values were compared to the  $\alpha_{corrected}$ .

Finally, a full G-DINA fit was considered at the model-level. A fully saturated model, or even a model with two-level or higher level interactions, is difficult to interpret, even though it may be substantively interesting. Therefore, model fit summaries such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIS), and the value of the loglikelihood (LL) were compared across all models to comment on the general appropriateness of this modeling framework for the substantive problem at hand.

#### 4.1.2.2 Context Skill and Skill Prevalence

To address the second question (RQ7), DINA models were fit to BLIS as well as MBLIS data from the pilot study (Chapter 1). Comparing model fits across data from both assessments allowed us to comment on whether the isomorphic instrument, MBLIS, provided comparable measurement of statistical literacy. Model level comparisons were not sufficient to achieve this goal. Therefore, estimated skill prevalence was compared.

Four DINA models were fit for this purpose. For each of the two datasets, one from BLIS and one from MBLIS, parameters of DINA models with and without the context skill were estimated. Skill prevalence is the estimated proportion of the sample which has achieved mastery over the given skill. This proportion was estimated for each skill under each of the four models. Standard errors for these proportions are required to be bootstrapped and were, therefore, not evaluated at this stage.

## 4.2 Results

### 4.2.1 Context Skill and Item Response Probabilities

#### 4.2.1.1 DINA models

First, we look at the AND  $\delta$  from the DINA models to consider whether the model which includes the context skill in the Q-matrix suggests different probabilities of responding to MBLIS items correctly. In Table 4.1, rows shaded in blue mark the anchor items from MBLIS plus an additional item - item 31. We note that the probabilities of answering correctly are barely affected based on the choice of the Q-matrix. However, the two highest differences across the probabilities are for anchor items - items which were unchanged across the two versions of the assessment and did not include a context. These were items 16 and 13, in descending order of absolute difference in probabilities.

Item	Without context skill		With context skill	
	AND $\delta$	SE	AND $\delta$	SE
1	0.3737	0.0399	0.3710	0.0394
2	0.2117	0.0505	0.2109	0.0502
3	0.3286	0.0483	0.3260	0.0480
4	0.2508	0.0339	0.2526	0.0338
5	0.3005	0.0477	0.2852	0.0455
6	0.5406	0.0564	0.5550	0.0566
7	0.4462	0.0456	0.4509	0.0453
8	-0.0526	0.0475	-0.0465	0.0473
9	0.1206	0.0461	0.1201	0.0459
10	0.4527	0.0456	0.4512	0.0455
11	0.0742	0.0466	0.0745	0.0464
12	0.1310	0.0500	0.1266	0.0496
13	0.1342	0.0482	0.1046	0.0661
14	-0.0227	0.0415	-0.0231	0.0412
15	0.5281	0.0456	0.5207	0.0451
16	0.1325	0.0453	0.2819	0.0624
17	0.4905	0.0451	0.5161	0.0461
18	-0.0036	0.0505	-0.0043	0.0504
19	0.4630	0.0439	0.4647	0.0437

Item	Without context skill		With context skill	
	AND $\delta$	SE	AND $\delta$	SE
20	0.3546	0.0474	0.3562	0.0472
21	-0.0949	0.0333	-0.0953	0.0329
22	0.2120	0.0452	0.2121	0.0452
23	0.1841	0.0504	0.1846	0.0503
24	0.3577	0.0418	0.3574	0.0418
25	0.3868	0.0402	0.3834	0.0405
26	0.2761	0.0486	0.2746	0.0485
27	0.1992	0.0492	0.1984	0.0492
28	0.4929	0.0364	0.4935	0.0364
29	0.6004	0.0435	0.5984	0.0437
30	0.6934	0.0410	0.6934	0.0411
31	0.2747	0.0299	0.2867	0.0310
32	0.4390	0.0449	0.4380	0.0449
33	0.3966	0.0400	0.3966	0.0399
34	0.4348	0.0496	0.4286	0.0499
35	-0.1698	0.0384	-0.1702	0.0382
36	0.4809	0.0387	0.4762	0.0385
37	0.4120	0.0423	0.4122	0.0423

Table 4.1: DINA estimates and standard errors

#### 4.2.1.2 A-CDM models

Next, we consider the main effect for the context skill under the additive CDM (A-CDM) model to investigate whether its effect on the probability of answering an item correctly, over and above the effect of all other required skills, has a small p-value. Table 4.2 shows the main effect estimate, its standard error, and the bonferroni adjusted p-value. It is important to note that since the anchor items and item 31 did not require familiarity with COVID-19 pandemic related contexts, this  $\delta$  was not estimated for those items. Eleven eligible items have a p-value below the cutoff point, and items 12 and 28 have p-values above but close to the cutoff point of 0.0017. For all these items, we consider that the probability of answering them correctly may be influenced by mastery (or the lack of it) of the context skill.

Item	Context main effect - delta	Context main effect - SE	p-value
1	0.3784	0.0683	0.0000
2	0.1180	0.0792	0.0681
3	0.4766	0.0695	0.0000
4	0.3095	0.0570	0.0000
5	0.2512	0.0524	0.0000
6	0.4047	0.0724	0.0000
7	0.1354	0.0618	0.0142
8	-0.0074	0.0771	0.4618
9	0.1201	0.0722	0.0481
10	0.2415	0.0695	0.0003
11	-0.0645	0.0768	0.2005
12	0.2076	0.0719	0.0019
13	NA	NA	NA
14	-0.0148	0.0709	0.4173
15	0.2506	0.0671	0.0001
16	NA	NA	NA
17	NA	NA	NA
18	-0.1038	0.0832	0.1061
19	0.2598	0.0665	0.0000
20	-0.0896	0.0652	0.0847
21	0.0012	0.0545	0.4912
22	0.1051	0.0723	0.0730
23	NA	NA	NA
24	NA	NA	NA
25	0.1292	0.0865	0.0676
26	0.2428	0.0723	0.0004
27	0.0154	0.0744	0.4180
28	0.2326	0.0797	0.0018
29	-0.0217	0.0763	0.3881
30	0.0015	0.0731	0.4918
31	NA	NA	NA
32	NA	NA	NA
33	0.2016	0.0797	0.0057
34	-0.0731	0.0793	0.1783

Item	Context main effect - delta	Context main effect - SE	p-value
35	-0.2069	0.0587	0.0002
36	0.3507	0.0708	0.0000
37	0.1706	0.0761	0.0125

Table 4.2: ACDM context skill main effect estimates and standard errors

### 4.2.1.3 Model comparisons

Finally, we consider model diagnostics for all models discussed above as well as the fully saturated G-DINA for MBLIS with the context skill in Table 4.3. Number of parameters estimated for each model are also listed.

	DINA model			
	w/o context	w context	ACDM w context	GDINA w context
AIC	27345.69	27597.71	27197.37	27172.34
BIC	28234.44	29052.42	29164.99	32164.35
LL	-13471.85	-13469.86	-13153.68	-12457.17
Parameters	201	329	445	1129

Table 4.3: Comparing various models for MBLIS

The model fits themselves do not look very different across the two models with Q-matrices that depend on whether the context skill was included in the Q-matrix or not. At the small scale of differences observed here, the DINA model which does not include the context skill has a smaller AIC and BIC, and would therefore be considered to have a better fit. However, towards the purpose of answering the first research question (RQ6), the probability of answering an item correctly seems to be impacted by the context skill when we assume an additive effect of each skill. This is not the case when a multiplicative effect is considered under the assumption that all requisite skills must be mastered in order to answer correctly.

## 4.2.2 Context Skill and Skill Prevalence

We look at skill prevalence with a focus on the mastery proportion of skills for all four models discussed in Section 4.1.2.2. Table 4.4 shows proportion of individuals who master each skill based on their responses to each of the two assessment instruments. Since these are not parameter estimates as per the model specification, and derived based on other parameters, standard errors are not available.

	BLIS		MBLIS	
	No context skill	With context skill	No context skill	With context skill
Skill 1	0.6483	0.6470	0.6098	0.7518
Skill 2	0.8298	0.8239	0.7391	0.8234
Skill 3	0.4465	0.4465	0.4643	0.4864
Skill 4	0.8035	0.8018	0.6515	0.6134
Skill 5	0.8837	0.9205	0.6847	0.9076
Skill 6	0.7340	0.7669	0.6729	0.8350
Skill 7	0.7671	0.7610	0.6762	0.6760
Skill 8	NA	0.9359	NA	0.7258

Table 4.4: Comparing skill prevalence across models

We must note that for BLIS, the original assessment instrument, the estimated prevalence is robust for the first seven skills, no matter whether the context skill is included in the Q-matrix or not. This indicates that the context skill does not affect the prevalence of other skills. However, the high prevalence of the context skill on BLIS is a critical reminder that the modeling framework may be detecting effects which may not be interpretable or must be interpreted post-hoc as is the case with all latent models. For MBLIS, which is the assessment of interest, skill prevalence estimates are different across the two model fits. If mastering the context skill is considered essential for being able to answer the 30 items correctly, the resulting model estimates higher prevalence of all skills except the visualization skill. However, comparing the first and the last column allows us to answer the second research question (RQ7) by observing that assuming that the current form of the Q-matrix and model specification are accurate, the two groups of students randomly determined to take either BLIS or MBLIS seem to have comparable prevalence of statistical skills considered to contribute to statistical literacy.



## 4.3 Discussion

This work demonstrates the feasibility as well as challenges of applying cognitive diagnostic modeling to constructs in statistics education, specifically statistical literacy, by developing an expert-specified Q-matrix that captures cognitive skills essential to correctly answering tasks contributing to the construct. Though the results are inconclusive, this work is a proof of concept that can encourage future work in this area. This work also underscores the argument that statistics education research is an inherently interdisciplinary endeavor [201] which must continue to bring together elements of cognitive theories, educational psychology, research on learning and teaching, as well as statistical methodology.

This study can be a starting point for future research focused on measuring statistical problem solving performance by explicitly considering the underlying cognitive skills. The measurement of the role of context in this problem solving can be further elicited by operationalizing the context in various different ways in an assessment. With the understanding that measurement models must, ideally, be used not only to analyze data from existing assessments but also to carefully design future assessments, similar work can contribute to designing an assessment of statistical literacy which can specifically measure the role of context familiarity.

This study analyzed data using available CDMs. However, extensions to this framework may be considered wherein a combination of AND and OR requirements can be specified to capture the skills which must be mastered for each item. For example, it is conceivable that items on the assessment of contextualized statistical literacy could be answered correctly if a test-taker has all requisite statistical skills (AND), and familiarity with the context could affect the probability of success but is not essential to success (OR). Alternatively, a framework in which mastery of one or more of the skills can be specified along a continuum rather than as a dichotomous possibility could also be beneficial.

### 4.3.1 Limitations

Findings from this modeling exercise are limited by the assumption that the Q-matrix in Appendix D accurately captures the cognitive skills essential to the measurement of statistical literacy as constructed by the assessment instruments. However, a CDM model does not interpret the substantive meaning of an additional column such as the context skill. This must be accounted for when interpreting results. Additionally, retrofitting data from an assessment which is not developed for the specific diagnostic purpose nor

based on the diagnostic analysis of the type conducted in retrofitting must be treated with caution. Diagnostic modeling is most valuable in situations where a Q-matrix is constructed a-priori, a test is developed to measure those particular skills, and then feedback is provided to test-takers based on analyzing those data. Additionally, even though the context skill column assumes that each isomorphic item based on a topic pertaining to the COVID-19 shares a context, it may be possible to argue that each different topic itself forms a different context and therefore, should be treated as such.

Additionally, as listed in Table 4.3, the sheer number of parameters in a CDM could potentially reduce the reliability of the estimates.

### **4.3.2 Implications for research**

As discussed in Section 4.1.1, the definition of statistical literacy used in this work states: “Statistical literacy is the ability to read, understand, and communicate statistical information.” Future work could incorporate a lexical score for reading load or a reading comprehension score for each item stem and answer options to incorporate this dimension of statistical literacy. Alternative framing of a Q-matrix for measuring statistical literacy using other assessments will also be beneficial for considering the effect of context in measurement and assessment of statistical literacy. However, the most critical research opportunity is a closer look at the cognitive skills essential for a statistically literate citizen.

This work specified the effect of context skill on probability of answering an item correctly as either contributing to a full-interaction between all required skills or as an additive coefficient. CDMs are specified in such a way that this coefficient is different for each item. However, since the context is common across all items, a substantively interesting model would be one where the context coefficient can be constrained to be equal for all items where context skill is needed.

# Chapter 5 |

## Conclusion

This work makes several methodological contributions despite some of the inconclusive results. We demonstrate that a carefully designed isomorphic assessment can allow for reliable assessment of statistical literacy in specific contexts. This assessment indicates that context matters because a year into the COVID-19 pandemic (as of April 2021), students completing a semester of college-level introductory statistics scored lower on a pandemic-specific assessment of statistical literacy as compared to another version with a variety of non-pandemic contexts. We illustrate through an example that a well-designed randomized experiment can allow for drawing causal conclusions about the effects of a curricular treatment in a statistics education research study. This study also underscores the value of using research-based assessments to measure outcomes of interest. Finally, we apply a measurement model to statistics education, developing a schema for the cognitive skills underlying statistical literacy. The consistent theme of exploring the role of relevant context in statistically literate behavior, through its assessment, measurement, and improvement, also contributes to substantive research aimed at understanding the role of context in not only statistical literacy, but also statistics education at-large.

Analyses of the pilot study in Chapter 5.1 inform two distinct questions at hand and what we learn from one informs the other. On the topic of isomorphic assessments, we set out to investigate whether the M-BLIS measures the same underlying constructs as BLIS and in the same way, or not. We can conclude that a carefully constructed isomorphic assessment can measure the same underlying constructs while exposing the test taker to statistical literacy concepts through the lens of a variety of application areas. Future work should look at various characteristics of the items as well as test-takers to further understand differences in assessment performances. The second question of interest to us was a comparison of student performance. These analyses were conducted assuming that scores on BLIS and M-BLIS are equatable under the internal-anchor design [193].

Various CTT-based analyses using multiple linear regression indicated that assessment type was an important predictor of total score no matter which other characteristics were included and whether the model includes any interactions or not.

Results discussed in Chapter 3 allowed us to take a closer look at the implicit assumption that students develop an ability to apply learnings from the course to important contexts encountered in their lives outside the classroom as well as contexts impacting the world around them. This assumes knowledge transfer [65–67] and warrants further study to investigate the nature and the extent of this transfer. Further, the statistics education community has been striving to bring real data into the classrooms. However, real data need not always be relevant. Connecting contextualized statistical literacy to the types of data included in the course materials provided us an insight into whether this ability can be further improved, informing future statistics educators.

Finally, Chapter 4 was a first step towards future research focused on measuring statistical problem solving performance by explicitly considering the underlying cognitive skills. This work can be extended to conduct foundational research that identifies cognitive skills required for statistical problem solving and operationalizes them through assessment design and analyses of item response data using appropriate measurement models. Specifically, this work can offer an insight into the role of context through the use of state-of-the-art statistical methodology.

## 5.1 Limitations

This research is not without its limitations and we discuss those in this section. When developing MBLIS, the less portable items on BLIS required either raw quantitative data, data from a randomized experiment, or data that led to visualizations with peculiar characteristics such as strong right skewness. Concessions were made in case of three items where for one item the parameter of interest was switched from mean to proportion, and an observational study was discussed instead of a randomized experiment in one other. As seen in the example in Table 2.2, reverse skewness was accepted for one item. However, the lack of open availability of raw datasets is a hurdle that will need to be addressed more systematically in creating future isomorphs. Additionally, balancing the competing goals of maximizing engagement and minimizing emotional impact lead to the inclusion of some topics which may not be most relevant to the lives of our target population for the study - college students, in this case - and exclusion of some topics which may be directly related to them. For example, one of the modified items

referred to pre- and during pandemic performance of elementary school students on standardized tests. This issue is confounded by the expectations of the ‘college student’ audience which is typical to an educational research study, though that may not need to be the case for the general purpose of the research. The choice of the test population can bias the choice of relevant contexts. The item in Table 2.5 was a subject of lengthy discussions, some of which included the expert reviewers. The implicit assumption of a coin being unbiased and our intuition about 50% of them landing on heads benefitted the original item. However, upon deliberation, it was agreed that it is extremely hard to find other phenomena which have an unconditional 0.5 probability of occurrence which is understood intuitively, and therefore the substantial change in wording was included. The original item was an interesting case because students are assumed to be so familiar with fair coins that the frequency of their ‘encounters’ with the context might actually outweigh the other dimensions of engagement/relevance we are seeking in this study. Authors must also acknowledge that even though we use anchor items to compare the two sets of respondents at baseline, we have to account for possible ordering effect. These identical items could function differently across BLIS and M-BLIS, especially since they may appear out-of-context on an assessment based entirely on one specific topic - the COVID-19 pandemic. Finally, survey questions were asked at the end of the assessment. Therefore, we didn’t expect that students’ performance on the assessment would have been affected by these. However, responses to the survey questions may have contained some cognitive bias based on whether they had just seen an entire assessment based on COVID-19 or not.

We suggest caution in concluding the lack of causal effect of the treatment in Chapter 3 due to three aspects of the study design which may have limited the effects observed in the study. First, this study was conducted during first fully in-person semester following the COVID-19 pandemic. Resultantly, students were not required to attend the lab sessions during which they could receive support from the TA and LA, as well interact with students within their section. This could have led to interference due to students working with fellow classmates from other lab sections. Second, since only a subset of the lab activities was modified for the treatment, that may have limited the effect size. Finally, it is worth noting that the modification of lab activities to include relevant contexts was only a small part of the curriculum. All other components of the course including the unmodified labs, lecture materials, common homework assignments, and exams were unchanged. Over the course of past several years, the instructional team for this course had undertaken concerted efforts to make this course relatable and relevant

to students. This could have potentially limited the size of the treatment effect. On the outcome side, [17] discussed the validity of BLIS as a mid-term and end-term assessment. M-BLIS was only deployed as a post-test in . Therefore, the use of these assessments as a pre-test in the present study should be treated with care. However, the design of BLIS is such that none of the items are invalid as a pre-test. Finally, the final dataset in this study was approximately 10% of the enrollment in the course. Even though Missingness At Random is assumed for the attrition and subgroup sizes within available data were comparable, it is difficult to estimate whether larger retention could have lead to different results. We also acknowledge the limited generalizability of these results since the experiment was conducted at a single institution and in a single course. However, this is typical of educational studies and future work will be aimed at multi-site studies.

Finally, findings from applying cognitive diagnostic modeling to data from assessments of statistical literacy are limited by the assumption that the Q-matrix in Appendix D accurately captures the cognitive skills essential to the measurement of statistical literacy as constructed by the assessment instruments. However, a CDM model does not interpret the substantive meaning of an additional column such as the context skill. This must be accounted for when interpreting results. Additionally, retrofitting data from an assessment which is not developed for the specific diagnostic purpose nor based on the diagnostic analysis of the type conducted in retrofitting must be treated with caution. Diagnostic modeling is most valuable in situations where a Q-matrix is constructed a-priori, a test is developed to measure those particular skills, and then feedback is provided to test-takers based on analyzing those data. Additionally, even though the context skill column assumes that each isomorphic item based on a topic pertaining to the COVID-19 shares a context, it may be possible to argue that each different topic itself forms a different context and therefore, should be treated as such.

## **5.2 Implication for Future Work**

### **5.2.1 Implications for research**

Since BLIS and MBLIS instruments were observed to function comparably, we argue that isomorphic assessment can be created to assess statistical literacy in various pertinent contexts. Even though it may be quite tedious create them, these instruments can be invaluable tools in getting respondents to consider statistics through a contextual lens that is relevant, and continue to measure how curricular strategies may affect literacy

levels. Therefore, future research can be directed towards two purposes. 1) measurement of statistical literacy in various disciplinary or societal contexts using isomorphs of BLIS, and 2) using these isomorphic versions to assess performance of experimental curricular or pedagogical strategies. However, additional work exploring the transfer and cognitive processes behind statistical problem solving will also be essential to our understanding the role of contexts.

The pilot study was intended to study psychometric properties of M-BLIS in comparison with BLIS to determine whether the BLIS and M-BLIS are psychometrically isomorphic, and whether they measure the same constructs even when the context is changed. To draw reliable conclusions, it was essential that we have the ability to compare results from our study to the field test conducted during the development of the original assessment. To achieve this, it was important to ensure that the BLIS items remained identical to that test, and therefore, M-BLIS was based on that version. At no point did we change any details in the original assessment in an effort to ensure comparability across the original work [17] and our pilot study. Resultantly, the results from this paper are specific to one definition and assessment of statistical literacy. Future research should study the role of contexts using other assessment instruments.

Differential student performance on BLIS and M-BLIS with a low p-value on inferential results indicates that the context in which a statistical question is posed affects assessment responses. Our respondents made sense of statistical questions differently based on whether the context behind the numbers was relevant to them or not. In reference to the discussion in Section 2.2.1 regarding sensitive contexts, this finding also has implications for teaching practices. If additional research finds that the sensitivity of the topic may have contributed to the lower scores on M-BLIS, an argument can be made to favor inclusion of such topics on curricular materials instead of including them in grade-affecting assessments [194].

From a context point of view, two things are worth noting. First, as discussed in Section 2.3.1, some of the BLIS items pertaining to college students saw better performance even though the examples were realistic. This may suggest that relevance itself may be hypercontextualized for different subgroups. Secondly, it was interesting to note in Table 2.11 that there was a certain percentage of students who, no matter which assessment they took, indicated after completing the assessment that they had engaged with the COVID-19 pandemic by seeking out information, believed it was relevant to their lives, yet would not be interested in gaining data-driven insights into the pandemic. Granted, this study ran about 13 months into the pandemic and there may have been

pandemic fatigue. However, this was at the end of a semester during which they had taken an introductory statistics class, making this an interesting phenomena warranting further investigation.

Future researchers interested in investigating the impact of teaching through contexts relevant to students must consider several important aspects. First, the list of topics this intervention worked with was pre-specified by the researcher. It would be worthwhile to build this list using open-ended responses from students themselves to include topics students report as engaging and relevant. Second, student engagement with topics can differ based on the component of the course in which they are incorporated. Future work should consider the possible differential effect of modifying other parts of the curriculum and assessments. Finally, contextualized statistical literacy is one possible measure of someone's ability to make sense of statistics pertaining to relevant contexts. However, it is possible to conceive of other ways of assessing the transfer of classroom learning to topics relevant to individuals. With the larger goal of promoting a statistically literate citizenry, other methods of measuring the abilities deserve merit as well.

On the methodological side, the curricular experiment established the feasibility of and a framework for conducting studies which can allow for causal claims to be made in educational research. Such a pursuit is not without its challenges such as infeasibility of randomized assignment at an individual-level, buy-in from the administration, consent procedures required by individual institutional review boards, and ensuring equal learning opportunity to the experimental and control group. However, more research-based causal claims regarding the effects of instructional ideas on important learning outcomes are critical to making evidence-based teaching decisions. Therefore, future statistics education research should consider conducting more randomized experiments or carefully designed observational studies and drawing careful causal conclusions based on those.

Finally, as discussed in Section 4.1.1, the definition of statistical literacy used in this work states: "Statistical literacy is the ability to read, understand, and communicate statistical information." Future work on applying CDM to statistical literacy could incorporate a lexical score for reading load or a reading comprehension score for each item stem and answer options to incorporate this dimension of statistical literacy. Alternative framing of a Q-matrix for measuring statistical literacy using other assessments will also be beneficial for considering the effect of context in measurement and assessment of statistical literacy. However, the most critical research opportunity is a closer look at the cognitive skills essential for a statistically literate citizen.

This work specified the effect of context skill on probability of answering an item



correctly as either contributing to a full-interaction between all required skills or as an additive coefficient. CDMs are specified in such a way that this coefficient is different for each item. However, since the context is common across all items, a substantively interesting model would be one where the context coefficient can be constrained to be equal for all items where context skill is needed.

The measurement of the role of context in this problem solving can be further elicited by operationalizing the context in various different ways in an assessment. With the understanding that measurement models must, ideally, be used not only to analyze data from existing assessments but also to carefully design future assessments, similar work can contribute to designing an assessment of statistical literacy which can specifically measure the role of context familiarity.

This study analyzed data using available CDMs. However, extensions to this framework may be considered wherein a combination of AND and OR requirements can be specified to capture the skills which must be mastered for each item. For example, it is conceivable that items on the assessment of contextualized statistical literacy could be answered correctly if a test-taker has all requisite statistical skills (AND), and familiarity with the context could affect the probability of success but is not essential to success (OR). Alternatively, a framework in which mastery of one or more of the skills can be specified along a continuum rather than as a dichotomous possibility could also be beneficial.

Lastly, but most importantly, all of the methodological advances discussed here should be applied more broadly to topics in statistics education research.

### **5.2.2 Implications for teaching**

As the substantive focus of this research is statistics education, motivated by the goal of educating students more effectively, its implications towards classroom decision are critical to consider. Chapter 5.1 comments on the negotiation vis-a-vis sensitive relevant contexts and the possible advantage of including them on curricular material as opposed to assessment items. Inclusion of such context in teaching materials can have an added advantage that no matter the statistical literacy outcomes, the course materials encourage students to consider topics relevant to the society from a statistical point of view. However, additional considerations such as attitudes towards statistics and engagement with the subject are also equally important. If future studies lead to findings consistent with this work, contexts in a statistics classrooms should be chosen to optimize outcomes other than statistical literacy.

The central goal of applying CDM to assessment data is improved ability to provide feedback to students. Even though we do not report or discuss person-level parameter estimates in Chapter 4, the modeling framework does provide them. An important extension of this work would be to consider how it can be directed towards developing a feedback mechanism that can offer instructors with the tools required to work with individual students on understanding their own strengths and weaknesses, and developing strategies to harness the strengths and improve on the weaknesses.

## 5.3 Conclusion

[200] unequivocally asserted the mantra that not only should we assess what we value, but we must also teach what we wish to assess. We need to first outline the learning outcomes we value, create an assessment plan, and then, consequently, design curriculum to teach what we intend to assess. This dissertation discussed the value of context in statistically literate behavior, and an assessment, measurement, and instructional tool towards that purpose. As discussed in Chapter 1, statistics educators value students' ability to become statistically literate citizens who can make sense of data-driven information pertinent to their personal and professional lives. In Chapter 5.1, we discussed a proposal to assess such an ability - contextualized statistical literacy - using a research-based tool. Chapter 3 contributed to the literature on the final step of the cycle, designing teaching materials intended to improve statistical literacy and evaluate its effects on the intended outcome. Chapter 4 looked at the underlying measurement structure upon which all conclusions from the other two projects were based. For a statistically literate individual, the ability to marry one's understanding of statistical constructs and the context-at-hand is assumed. In fact, as there is no statistics without context [35], statistical literacy is also inherently contextualized. Parallel to the discussion in [72] in the context of mathematics education, statistics education, too, is a way for us to develop citizens who can make sense of quantitative information in contexts that matter to them. This work will support future steps in the methodological research essential towards this goal.

# Appendix A | Assessment Instrument - MBLIS

This appendix includes the blueprint for and final copy of MBLIS or Modified BLIS measuring contextualized statistical literacy. Blueprint and final copy of the Basic Literacy in Statistics (BLIS) assessment instrument is presented in [17] and available upon request from the original author.

## A.1 Blueprint for MBLIS

The original blueprint drawn and discussed by [17] includes the Topic and the Learning Outcome for each item. For the purpose of this presentation, we re-create the original blueprint which includes the learning outcomes before presenting the extended blueprint created for MBLIS without the learning outcomes.

Topic	Item	Learning Outcome
Data production	1	Understanding of the difference between a sample and population
	2	Understanding that randomness cannot be outguessed in the short term but patterns can be observed over the long term
	3	Understanding that statistics computed from random samples tend to be centered at the parameters
	4	Ability to determine what type of study was conducted
	5	Ability to determine if the variable is quantitative or categorical
	6	Ability to determine if a variable is an explanatory variable or a response variable
	7	Understanding of the difference between a statistic and parameter
	8	Understanding that statistics vary from sample to sample
	9	Ability to describe and interpret a dotplot
Graphs	10	Ability to describe and interpret the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data
	11	Understanding the importance of creating graphs prior to analyzing data
	12	Ability to interpret a probability in the context of the data
	13	Ability to interpret a mean in the context of the data
Descriptive statistics	14	Understand how a mean is affected by skewness or outliers
	15	Ability to interpret a standard deviation in the context of the data
	16	Understanding of the properties of standard deviation
Empirical sampling distributions	17	Understanding of what an empirical sampling distribution represents
	18	Understanding that an empirical sampling distribution shows how sample statistics tend to vary

Topic	Item	Learning Outcome
Confidence intervals	19	Understanding that simulated statistics in the tails of a sampling distribution are not plausible estimates of a population parameter
	20	Understanding that a confidence interval provides plausible values of the population parameter
	21	Understanding that a confidence interval for a proportion is centered at the sample statistic
	22	Understanding of how the confidence level affects the width of a confidence interval
Randomization distributions	23	Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis
	24	Understanding of how sample size affects the standard error
	25	Understanding that a randomization distribution tends to be centered at the hypothesized null value
	26	Ability to estimate a p-value using a randomization distribution
Hypothesis tests	27	Understanding of the logic of a hypothesis test
	28	Understanding of the purpose of a hypothesis test
	29	Ability to determine a null and alternative hypothesis statement based on a research question
	30	Ability to determine a null and alternative hypothesis statement based on a research question
	31	Ability to determine statistical significance based on a p-value
	32	Understanding that errors can occur in hypothesis testing
	33	Understanding of how a significance level is used to make decisions

Topic	Item	Learning Outcome
Scope of conclusion	34	Understanding that only an experimental design with random assignment can support causal inference
	35	Understanding of the factors that allow a sample of data to be generalized to the population
Regression and correlation	36	Ability to match a scatterplot to a verbal description of a bivariate relationship
	37	Ability to use a least-squares regression equation to make a prediction

Table A.1: Original BLIS blueprint [17]

Item	Testlet	BLIS example type - GAISE	Variable type(s)	Actual data need	Analysis need	Random sample	Randomized assignment
1		Real from real study	Categorical	Yes	No	Yes	
2		Realistic	Categorical	No	No		
3		Real	Quantitative	Maybe	Yes	Yes	
4		Real	2 Categorical	Maybe	No		Yes
5	1	Realistic	Categorical	No	No		
6	1	Realistic	Categorical	No	No		
7		Real from real study	Categorical	Yes	No	Yes	
8		Realistic	2 Quantitative	Maybe	Maybe	Yes	
9		Realistic	Quantitative	Maybe	Yes		
10		Realistic	Quantitative	Maybe	Yes		
11		Real	1 Categorical, 1 Quantitative	No	No		Yes
12		Real from real study	Categorical	Yes	No		
13		Real from real study	Quantitative	Maybe	Maybe	Yes	
14		Naked	Quantitative	Maybe	Yes		
15		Realistic	Quantitative	Maybe	Maybe		
16		Realistic	Quantitative	Maybe	Maybe		
17		Realistic	Quantitative	Maybe	Yes	Unspecified	
18	2	Real from real study	Quantitative	Yes	Yes	Yes	
19	2	Real from real study	Quantitative	Yes	Yes	Yes	
20		Real from real study	Categorical	Yes	No	Yes	
21		Real from real study	Categorical	Yes	No	Yes	
22		Realistic	Quantitative	No	Maybe	Yes	

Item	Testlet	BLIS example type - GAISE	Variable type(s)	Actual data need	Analysis need	Random sample	Randomized assignment
23	3	Real from real study	1 Categorical, 1 Quantitative	Yes	Yes		Yes
24	3	Real from real study	1 Categorical, 1 Quantitative	Yes	Yes		Yes
25	4	Real from real study	1 Categorical, 1 Quantitative	Yes	Yes		Yes
26	4	Real from real study	1 Categorical, 1 Quantitative	Yes	Yes		Yes
27		Realistic	Categorical	No	No		
28		Real from real study	1 Categorical, 1 Quantitative	Maybe	No		
29	5	Real from real study	1 Categorical, 1 Quantitative	Yes	No		
30	5	Real from real study	1 Categorical, 1 Quantitative	Yes	No		
31		Realistic					
32		Real from real study	1 Categorical, 1 Quantitative	Yes	No		
33		Real from real study	Categorical	Yes	Maybe		
34		Real	1 Categorical, 1 Quantitative				
35		Real from real study	Categorical	Maybe	No	Yes	
36		Real	2 Quantitative	Maybe	Yes		
37		Real from real study	2 Quantitative	Maybe	Yes		

Table A.2: Extended blueprint for MBLIS



## A.2 MBLIS instrument

### Item 1:

**Item stem:** The Pew Research Center surveyed a nationally representative group of 12,648 U.S. adults in November 2020. Of these adults, 62% said they would be uncomfortable being among the first to get the vaccine for COVID-19. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population.

Source: Pew Research Center.

- The population is all U.S. adults in November 2020. The sample is the 62% of U.S. adults who said they would be uncomfortable being among the first to get the vaccine for COVID-19.
- The population is the 12,648 U.S. adults surveyed. The sample is all U.S. adults in November 2020.
- The population is all U.S. adults in November 2020. The sample is the 12,648 U.S. adults surveyed.

---

### Item 2:

**Item stem:** Penn State University administrators surveyed all undergraduate students to capture feedback from the entire student body on several issues. As a result, they learned that 86% of all students planned to return in fall 2020. Despite knowing the proportion for all Penn State students as a whole, several instructors surveyed their own classes in order to be sensitive to the views of their students. One instructor had a class with 50 students and another instructor had a class with 100 students. Assuming both classes were representative of the entire student body at Penn State, which instructor was more likely to find that 84% to 88% of their students would plan to return in fall 2020?

Source: Adapted from Penn State News.

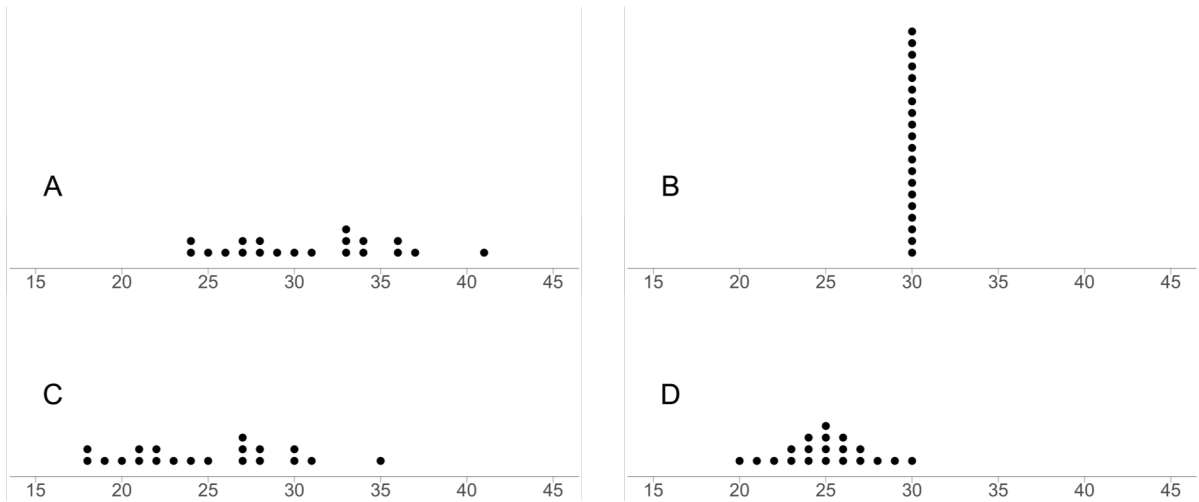
- The instructor who surveyed 50 students because the percent that planned to return was less likely to be exactly 86%.
- The instructor who surveyed 100 students because that instructor had more chances to survey a student who planned to return.

- The instructor who surveyed 100 students because the more students that were surveyed would have increased the chance of approaching a result of 86% planning to return.
- Neither instructor was more likely because student responses were random and therefore you could not predict the survey responses.

**Item 3:**

**Item stem:** Researchers learned that college students were getting 30 more minutes of sleep, on average, during weekdays of the first COVID-19 lockdown. To determine whether this finding could be replicated, additional random samples of 25 college students were taken and the average additional weekday sleep in each sample was recorded. Assuming that nothing was wrong with the initial study, which of the following graphs is the most plausible for the average additional sleep in each of the 20 samples?

Source: Research article.



- Graph A
- Graph B
- Graph C
- Graph D

**Item 4:**

**Item stem:** Johnson and Johnson wanted to determine if their proposed vaccine reduces the chance of developing mild COVID-19. Researchers at the company recruited 20,000 individuals into the Phase III clinical trial. Half (10,000) of the individuals were randomly assigned to receive the actual vaccine dose and the other half to receive a placebo. Then after 14 days, the percentage of mild COVID-19 cases for the individuals who received the actual vaccine and for those who did not receive the actual vaccine were reported. What type of study did the scientists conduct?

Source: JnJ study protocol.

- Observational
- Experimental
- Survey

---

**Items 5 and 6 refer to the following situation:**

Researchers at a university gathered data on the COVID-19 experience of individuals in the U.S. One of the variables measured was sexual identity. These data were coded using the following method: 1 = straight or heterosexual, 2 = bisexual, 3 = asexual, 4 = pansexual, 5 = gay or lesbian, 6 = no label, 7 = undecided, and 8 = other label.

Source: Research article.

**Item 5:**

**Item stem:** What type of variable is this?

- Categorical
- Quantitative
- Continuous

**Item 6:**

**Item stem:** The researchers planned to see if sexual identity of an individual is a predictor of their COVID-19-related psychological distress. Identify the response variable in this study.

- Individuals in the US

- Sexual identity
  - Psychological distress
  - Average psychological distress
- 

**Item 7:**

**Item stem:** RAND Corporation surveyed a nationally representative sample of 2,387 U.S. adults during May - June 2020 to determine "What proportion of U.S. adults have delayed or forgone getting dental care or going to the dentist due to the COVID-19 pandemic?" For the sample, 1,115 adults answered yes and 1,272 adults answered no. Identify the statistic and parameter of interest.

Source: Research article.

- The statistic is the sample proportion of adults who answered yes ( $1115/2387 = .467$ ) and the parameter is the 2,387 US adults who took part in the survey.
  - The statistic is the 2,387 U.S. adults who took part in the survey and the parameter is all U.S. adults.
  - The statistic is the proportion of all U.S. adults who have delayed or forgone getting dental care or going to the dentist due to the COVID-19 pandemic and the parameter is the sample proportion of adults who answered yes ( $1115/2387 = .467$ ).
  - The statistic is the sample proportion of adults who answered yes ( $1115/2387 = .467$ ) and the parameter is the proportion of all U.S. adults who have delayed or forgone getting dental care or going to the dentist due to the COVID-19 pandemic.
- 

**Item 8:**

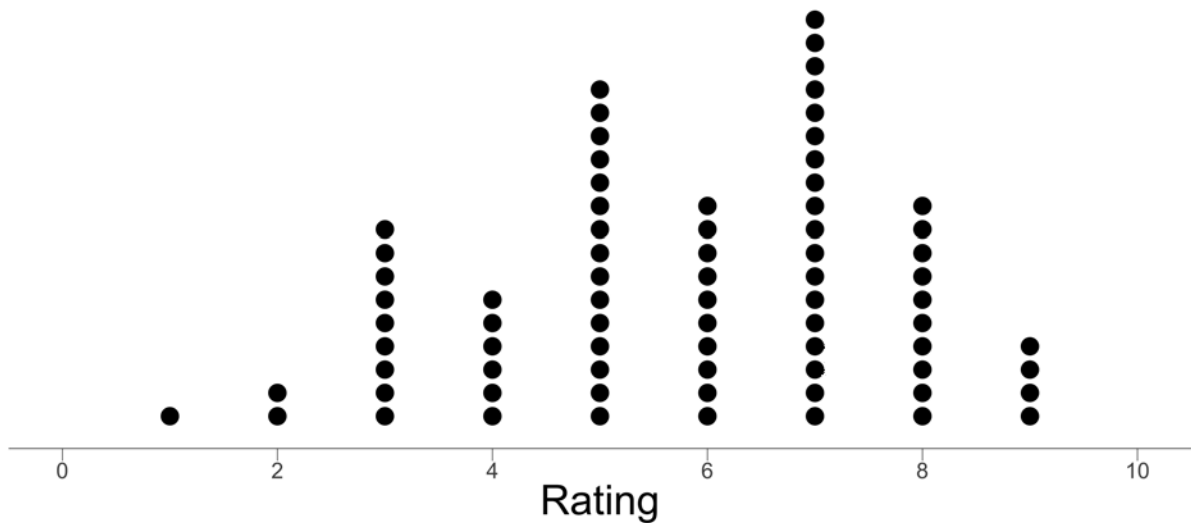
**Item stem:** In a study to investigate the impact of the COVID-19 pandemic on the global airline industry, Researcher A took a random sample of 25 days during the pandemic, and found that the mean number of flights (by large air carriers) canceled was 1765.6. In another study, Researcher B took a random sample of 25 days during the same period, and found that the mean number of flights (by large air carriers) canceled was 2278.72. What is the best explanation for why the samples taken by Researcher A and Researcher B did not produce the same mean?

Source: Kaggle.

- The sample means varied because they are small samples.
- The sample means varied because the samples were not representative of all days during the pandemic.
- The sample means varied because each sample is a different subset of the population.

**Item 9:**

**Item stem:** Researchers studied perceptions of Vietnamese citizens during the initial outbreak of the COVID-19 pandemic. One of the questions asked was "To what extent is the number of official news (regarding the pandemic) overwhelming on a scale of 1 to 10" where 1 = Least Overwhelming and 10 = Most Overwhelming. Below is the distribution of this variable for the 75 respondents with postgraduate education in the sample.



How should the researchers interpret Vietnamese citizens' perceptions regarding the official news during the initial outbreak of the COVID-19 pandemic?

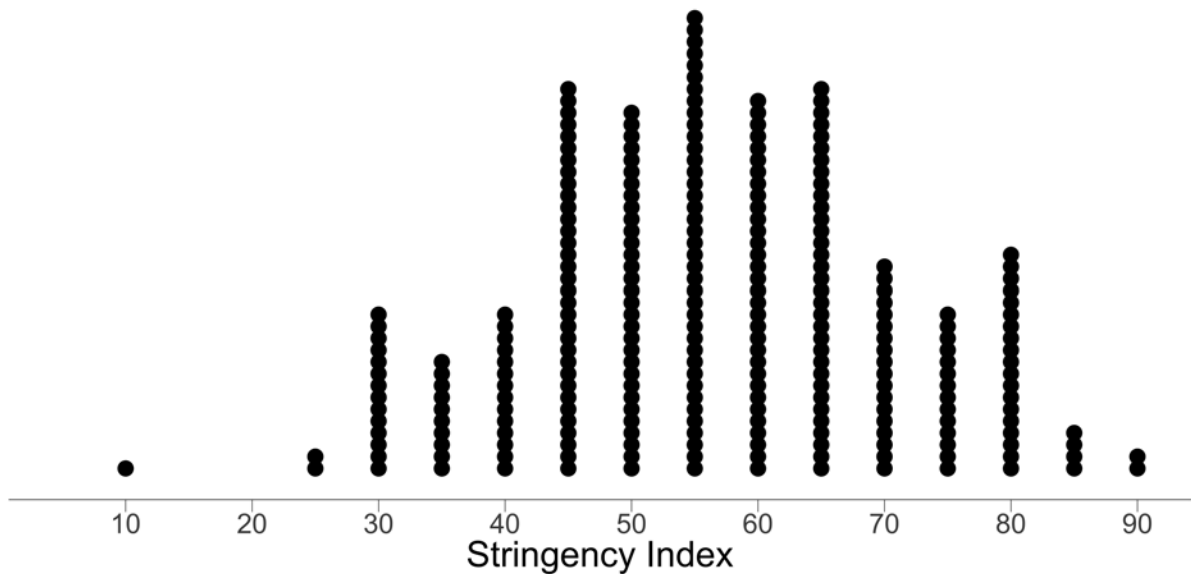
Source: Research article.

- A majority of the respondents felt overwhelmed by the number of official news although a few did not feel overwhelmed.
- A majority of the respondents rated their feeling of being overwhelmed as a 7 although some ratings were higher and some were lower.

- A majority of the respondents continued to access the official news even though they were overwhelmed by its number.

**Item 10:**

**Item stem:** Researchers at the University of Oxford have been collecting information on policy responses that governments have taken to respond to the pandemic. The following graph shows a distribution of the stringency index for a group of countries calculated on October 2, 2020. This index records the strictness of 'lockdown style' policies that primarily restrict people's behaviour.



Select the statement below that gives the most complete description of the graph in a way that demonstrates an understanding of how to statistically describe and interpret the distribution of a variable.

Source: Adapted from research data.

- The values go from 10 to 90, increasing in height to 55, then decreasing to 90. The most values are at 55. There is a gap between 10 and 25.
- The distribution is normal, with a mean of about 55 and a standard deviation of about 15.

- Many countries seem to have an index value of 55, but some countries have a higher value and some have a lower value. However, one country must have very lenient lockdown-style policies.
- The distribution of stringency indices is somewhat normal, with an outlier at 10. The typical index value is about 55 and standard deviation is about 15.

---

**Item 11:**

**Item stem:** Researchers at the National Institutes of Health (NIH) were interested in determining if taking Remdesivir was an effective treatment for adult COVID-19 patients with certain characteristics. An experiment was conducted with 1,062 participants. 541 of these participants were randomly assigned to receive Remdesivir and the others received a placebo. The number of days the participant took to recover was recorded. The researchers planned to conduct a hypothesis test to determine if there was a significant difference in the average number of days participants took to recover for the Remdesivir group and the placebo group. Which of the following is a reason why the researchers should create and examine graphs of the number of days participants took to recover before the hypothesis test is conducted?

Source: Research article.

- To decide what the null hypothesis and alternative hypothesis should be.
- To compute the average number of days participants took to recover in order to conduct a hypothesis test.
- To see if there are recognizable differences in the two groups to decide if a hypothesis test is necessary.

---

**Item 12:**

**Item stem:** Consider an individual fitting the following description.

- 20-year-old female,
- lives alone near a university campus,
- is exposed to an average of 10 people each week,
- has no underlying medical complications,

- is asymptomatic and unvaccinated,
- and follows CDC's guidance.

According to the "19andMe" tool developed by Mathematica, her probability of catching COVID-19 through community transmission in a week is .0024, as of March 30, 2021. What does the statistic, .0024, mean in the context of this calculation from Mathematica?

Source: Online calculator.

- For all individuals fitting the above description, approximately 0.24% will catch COVID-19 through community transmission at some point during the week.
- If you randomly selected an individual fitting the above description there is a 0.24% chance that they will catch COVID-19 through community transmission at some point during the week.
- In a random sample of 10,000 individuals fitting the above description, 24 of them will catch COVID-19 through community transmission at some point during the week.
- Both a and b are correct.

**Item 13:**

**Item stem:** According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is \$1,700. Which of the following is the best interpretation of the mean?

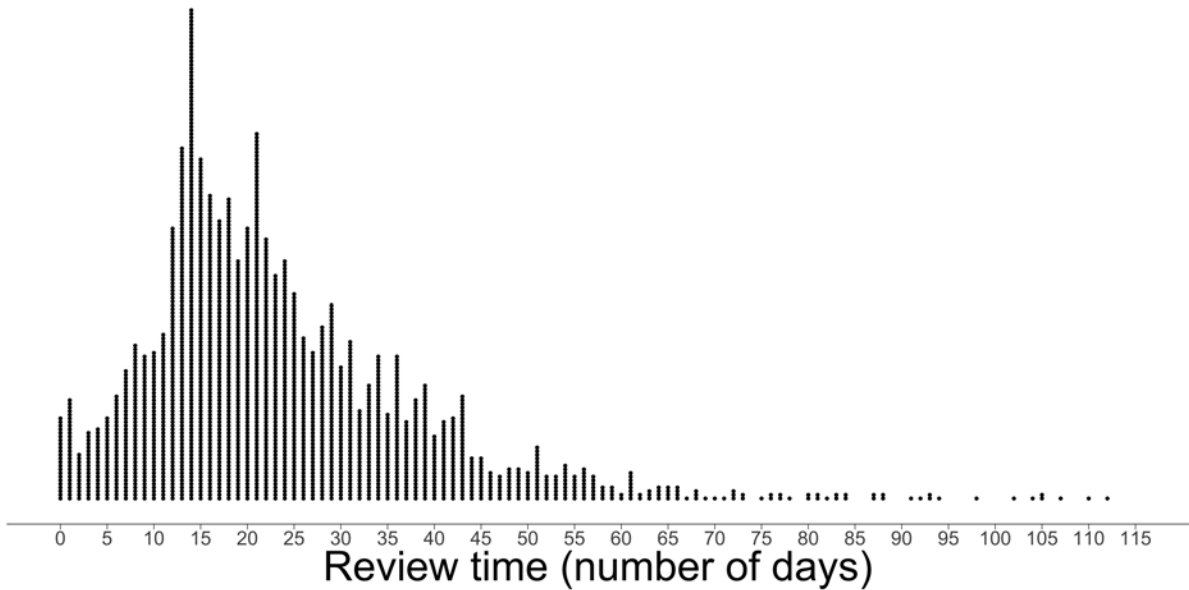
- For all dog owners in this sample, their average first-year costs for owning a large-sized dog is \$1,700.
- For all dog owners in the population, their average first-year costs for owning a large-sized dog is \$1,700.
- For all dog owners in this sample, about half were above \$1,700 and about half were below \$1,700.
- For most owners, the first-year costs for owning a large-sized dog is \$1,700.



---

**Item 14:**

**Item stem:** For scientific credibility, journal articles are reviewed by other scientists before publication. This process is called peer-review. Researchers collected data to study how the pandemic has affected the peer-review timelines for six Ecology journals. The plot below shows the distribution of number of days taken by all reviewers to review papers assigned to them.



A sample of 10 randomly selected papers will be taken from this population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?

Source: Research article.

- 0 to 10
- 10 to 20
- 20 to 30
- 40 to 50

---

**Item 15:**

**Item stem:** In the state of Pennsylvania, 147,469 Paycheck Protection Program (PPP) loans worth \$150,000 or less were issued from the beginning of the pandemic until August 8th, 2020. The standard deviation of the loan amounts was \$33,661. Which of the following gives the most suitable interpretation of this standard deviation?

Source: Treasury Department.

- All of the individual loan amounts are \$33,661 apart.
- The difference between the highest and lowest loan amount is \$33,661.
- The difference between the upper and lower quartile is \$33,661.
- A typical distance of a loan amount from the mean is \$33,661.

---

**Item 16:**

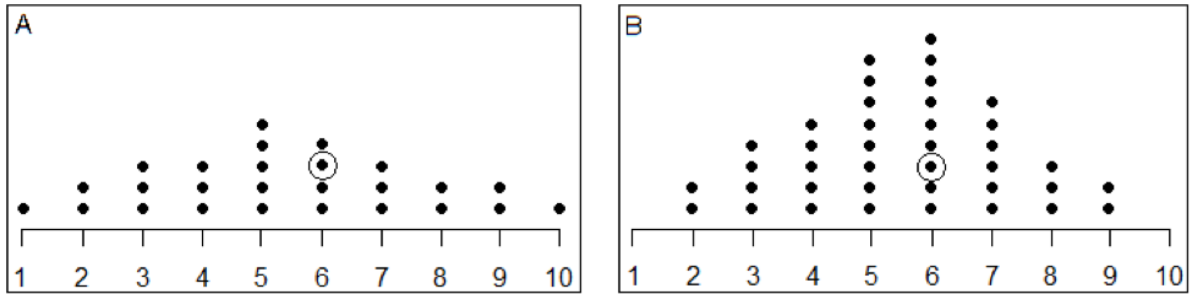
**Item stem:** A teacher gives a 15-item science test. For each item, a student receives one point for a correct answer; 0 points for no answer; and loses one point for an incorrect answer. Total test scores could range from -15 points to +15 points. The teacher computes the standard deviation of the test scores to be -2.30. What do we know?

- The standard deviation was calculated incorrectly.
- Most students scored below the mean.
- None of the above.

---

**Item 17:**

**Item stem:** Figure A represents the weights for a sample of 26 pebbles, each weighed to the nearest gram. Figure B represents the mean weights for 39 random samples of 3 pebbles each, with all mean weights rounded to the nearest gram. One value is circled in each distribution. Is there a difference between what is represented by the dot circled in A and the dot circled in B? Please select the best answer from the list below.



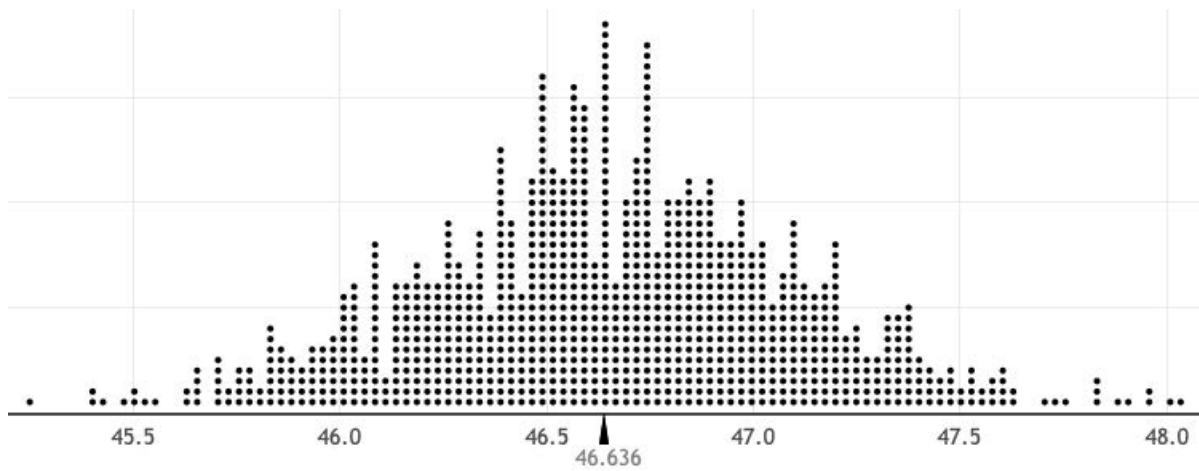
- No, in both Figure A and Figure B, the circled dot represents the same measurement, a weight of 6 grams.
- Yes, in Figure A there are only four dots with a weight of 6, but in Figure B there are nine dots with a weight of 6.
- Yes, the circled dot in Figure A is the weight for a single pebble, while the circled dot in Figure B represents the mean weight of 3 pebbles.

**Items 18 and 19 refer to the following situation:**

RAND collected data from 1,082 teachers during October 2020 and asked them "During the most recent full week, approximately how many hours did you work as part of your teaching position at your school?" The sample average number of hours worked was 46.6. An empirical sampling distribution was estimated by doing the following:

- From the original sample, 1,082 teachers were chosen randomly, with replacement.
- The mean was computed for the new sample and placed on the plot shown below.
- This was repeated 999 more times.

Below is the plot of the estimated empirical sampling distribution. (Note: This plot can be used as an estimate of what the sampling distribution would look like.)



**Figure A.1.** Mean Number of Hours

Source: RAND.

**Item 18:**

**Item stem:** Which of the following is the best description of the variability in the empirical sampling distribution?

- The mean number of hours teachers worked during October 2020 was 46.636.
- The variability in the mean number of hours teachers worked from sample to sample is quite small spanning from approximately 45 to 49.
- The variability in the number of hours worked from teacher to teacher is quite small, spanning from approximately 45 to 49.

**Item 19:**

**Item stem:** What values do you believe would be LESS plausible estimates of the population average number of hours worked by teachers if you wanted to estimate the population average with 95% confidence?

- Values approximately 47.3 and above because it is unlikely that teachers would work for so many hours in a week.
- Values below approximately 45 and values above approximately 48.5 because there are no dots that are that extreme.
- Values in the bottom 5% (below approximately 45.9) and values in the top 5% (above approximately 47.4).

- Values in the bottom 2.5% (below approximately 45.7) and values in the top 2.5% (above approximately 47.5).

---

**Item 20:**

**Item stem:** Gallup surveyed 3,759 randomly chosen U.S. adults during a week in February 2021. The sample percent of adults who visited a restaurant within the prior 24 hours was 24%. The 95% confidence interval was 20% to 28%. What is this interval attempting to estimate?

Source: Gallup.

- The average number of U.S. adults who visited a restaurant within the previous 24 hours during that week in February 2021.
- The percent of the 3,759 U.S. adults who visited a restaurant within the previous 24 hours during that week in February 2021.
- The percent of all U.S. adults who visited a restaurant within the previous 24 hours during that week in February 2021.
- For U.S. adults who visited a restaurant within the previous 24 hours during that week in February 2021, only 20% to 28% were comfortable visiting a restaurant.

---

**Item 21:**

**Item stem:** In a study of working mothers with children under 18 at home, researchers at the Pew Research Center randomly selected a sample and asked them if they have personally experienced being passed over for an important assignment during the pandemic because they were balancing work and parenting responsibilities. They calculated a 95% confidence interval for the percentage of mothers who said yes (10% to 12%). Which of the following statements is true about the center of the interval (11%)?

Source: Pew Research Center.

- We know that 11% of mothers in the sample have been passed over for an important assignment during the pandemic because they were balancing work and parenting responsibilities.
- We know that 11% of mothers in the population have been passed over for an important assignment during the pandemic because they were balancing work and parenting responsibilities.

- We can say with 95% confidence that 11% of mothers in the sample have been passed over for an important assignment during the pandemic because they were balancing work and parenting responsibilities.
- We can say with 95% confidence that 11% of mothers in the population have been passed over for an important assignment during the pandemic because they were balancing work and parenting responsibilities.

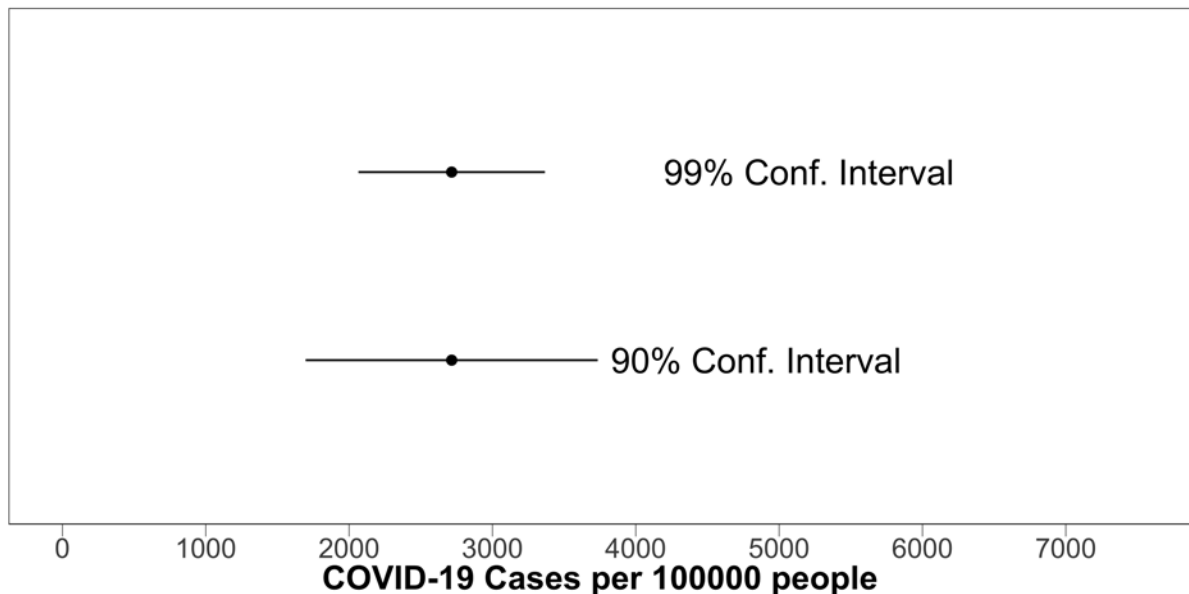
**Item 22:**

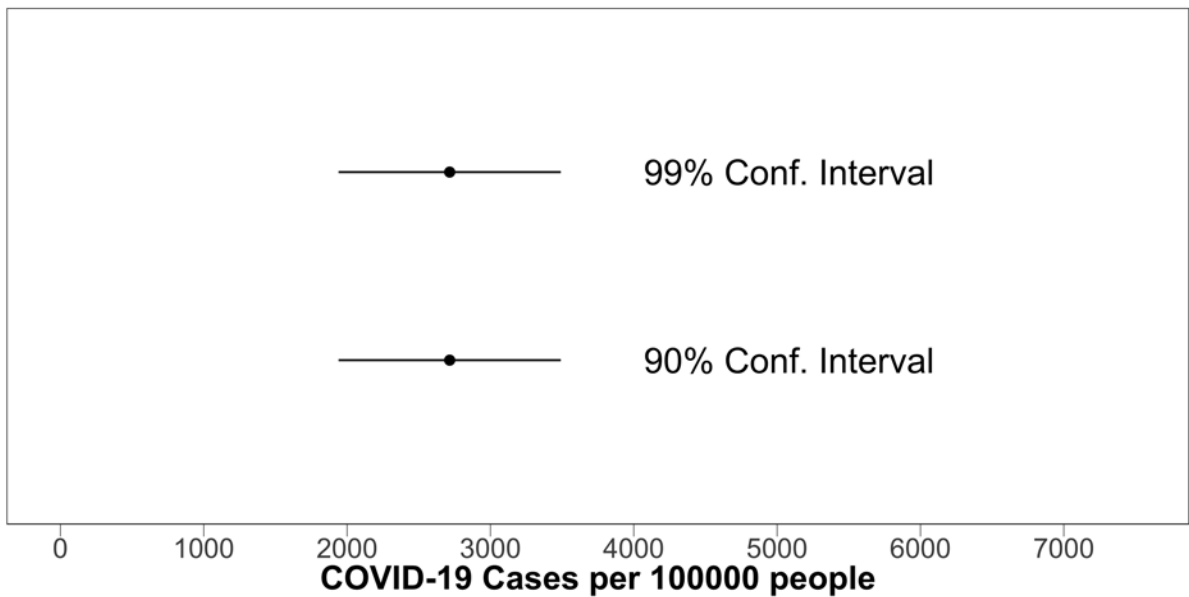
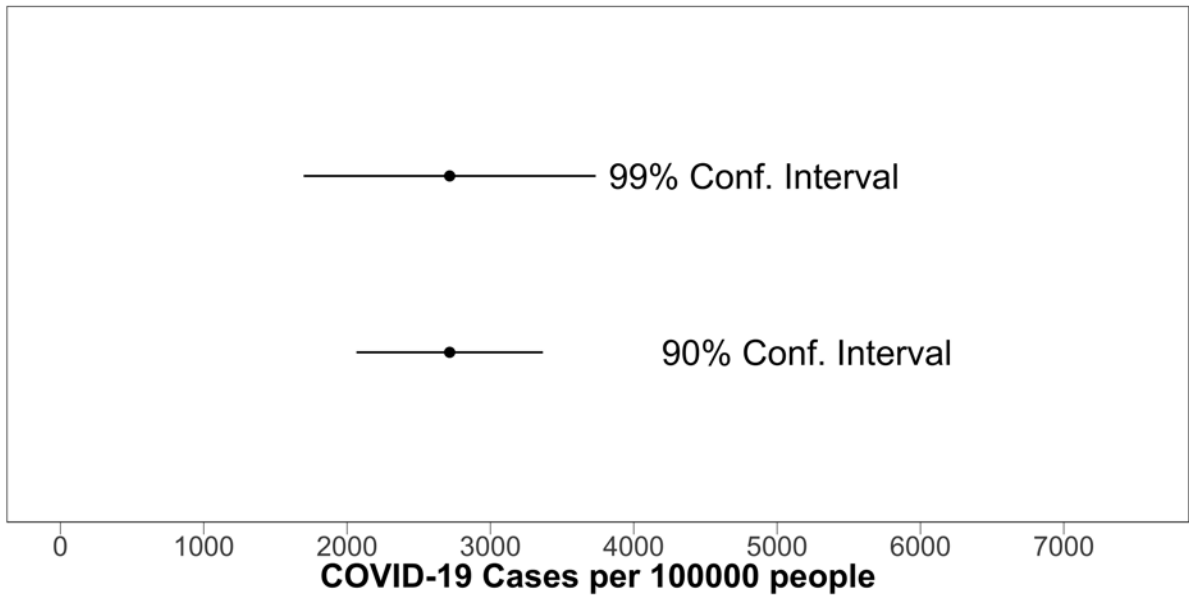
**Item stem:** The World Health Organization (WHO) maintains updated records of the number of COVID-19 cases in each country, territory and area.

On March 20th, 2021, we took a random sample of  $n = 50$  regions. We looked at the number of cases per 100,000 people in the population. A 99% confidence interval for the population mean and a 90% confidence interval for the population mean were constructed using this sample.

For the following options, a confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval. Which of the options would best represent how the two confidence intervals would compare to each other?

Source: WHO.





---

**Items 23 and 24 refer to the following situation:**

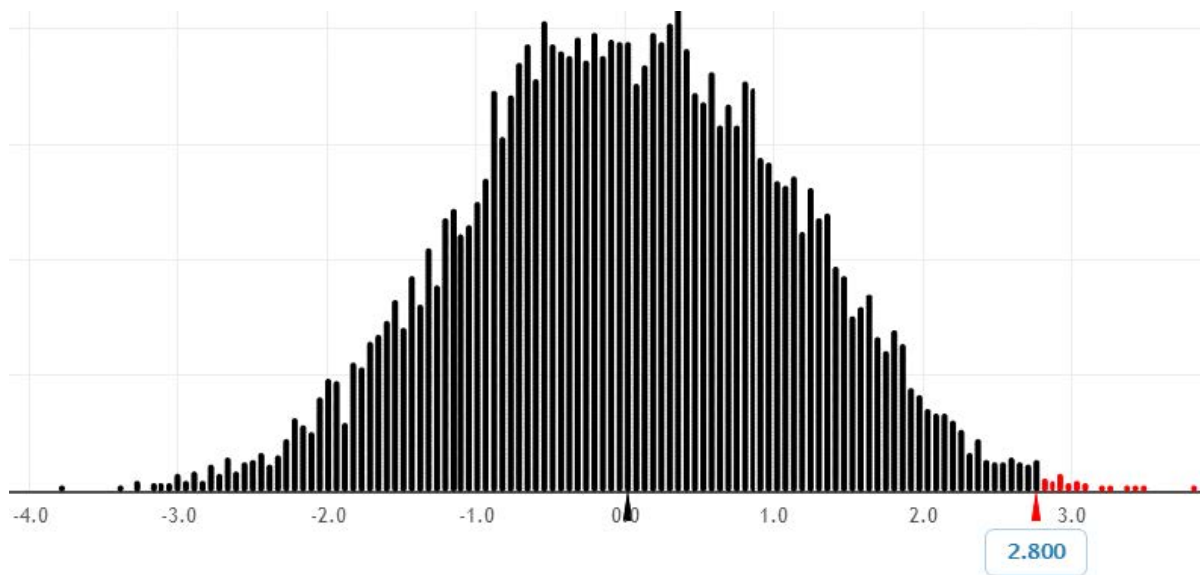
Are people able to recall words better after taking a nap or taking a caffeine pill? A randomized experiment was conducted with 24 participants. Participants were shown a list of words in the morning. In the afternoon, half of the participants were randomly assigned to take a nap and the other half took a caffeine pill. The response variable was

the number of words participants were able to recall 7 hours after being shown the list of words in the morning. The nap group recalled an average of 15.8 words and the caffeine group recalled an average of 13.0 words, with a mean difference of  $15.8 - 13.0 = 2.8$  words.

A randomization distribution was produced by doing the following:

- From the original sample, the 24 participants were re-randomized to the nap group (n=12) or caffeine group (n=12), without replacement.
- The mean difference in words recalled between the two re-randomized groups was computed [mean(nap group) - mean(caffeine group)] and placed on the plot shown below.
- This was repeated 999 more times.

Below is the plot of the randomization distribution. (Note: This plot can be used as an estimate of what the sampling distribution would look like if the null hypothesis is true.)



**Figure A.2.** (Mean Words Recalled for Nap Group) - (Mean Words Recalled for Caffeine Group)

**Item 23:**

**Item stem:** The null hypothesis is there is no difference in the true mean number of words recalled for the nap group and caffeine group. Looking at the observed sample



mean difference in number of words recalled between the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis?

- No, because the average of the re-randomized sample mean differences is equal to 0.
- No, because the proportion of re-randomized sample mean differences equal to or above 2.8 is very small.
- Yes, because the proportion of re-randomized sample mean differences equal to or above 2.8 is very small.
- Yes, because the observed result shows that the nap group remembered an average of 2.8 words more than the caffeine group.

**Item 24:**

**Item stem:** Suppose the sample size was doubled from 24 participants to 48 participants and the participants were still randomly assigned into two groups of equal size. How would you expect the standard error of the mean difference to change?

- Decrease, because with a larger sample size, there would be less variability in the re-randomized sample mean differences.
- Increase, because with a larger sample size, there is more opportunity for error.
- Stay about the same, because people are still being assigned to groups randomly.

---

**Items 25 and 26 refer to the following situation:**

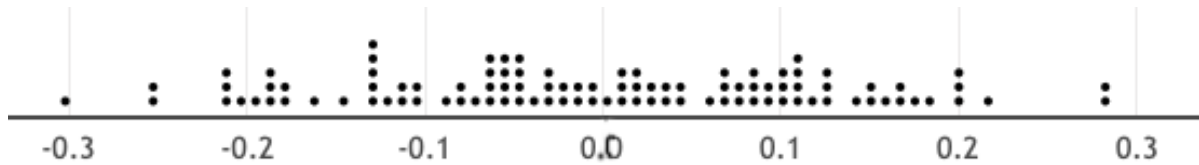
Researchers collected data from 89,305 people from around the world. Respondents were asked about their social distancing behavior during the COVID-19 pandemic. Of these people, 51,110 were co-habiting (including married), and 38,195 were single (including divorced). The average social distancing score was 83.07 for the co-habiting group and 80.41 for the single group. Note that a higher score implies that the individual was more socially distant. The difference in average social distancing score was  $83.07 - 80.41 = 2.66$ .

A randomization distribution was produced by doing the following:

- From the original sample, the 89,305 people were re-randomized to the cohabiting group ( $n=51,110$ ) or single group ( $n=38,195$ ), without replacement.

- The mean difference in social distancing score between the two rerandomized groups was computed [ $\text{mean}(\text{co-habiting}) - \text{mean}(\text{single})$ ] and placed on the plot shown below.
- This was repeated 99 more times.

Below is the plot of the randomization distribution for the 100 simulated mean differences. (Note: This plot can be used as an estimate of what the sampling distribution would look like if the null hypothesis is true.)



**Figure A.3.** (Mean Score for Co-habiting group) - (Mean Score for Single group)

Source: Research article.

**Item 25:**

**Item stem:** Why is the randomization distribution centered at 0?

- Because the randomization distribution was created under the assumption of a difference in mean social distancing score of 0.
- Because the people who had higher scores cancelled out the people who had lower scores resulting in a mean of 0.
- Because that was the original score that respondents started at for both groups.

**Item 26:**

**Item stem:** Researchers hypothesize that co-habiting is related to more social distancing than being single for people responding to this survey. Compute the approximate p-value for the observed difference in social distancing score of 0.25 based on the randomization distribution using the one-tailed test appropriate to the researchers' interest.

- .02
- .05

- .04

---

**Item 27:**

**Item stem:** The following situation models the logic of a hypothesis test. A procedure called RT-qPCR tests whether or not a wastewater sample is positive for the COVID-19 virus. The null hypothesis is that the sample is negative. The alternative hypothesis is that the sample is positive. The lab performs the test and decides to reject the null hypothesis. Which of the following statements is true?

Source: CDC.

- The sample is definitely positive and further action is needed.
- The sample is most likely positive, but it could be negative.
- The sample is definitely negative and no further action is needed.
- The sample is most likely negative, but it might be positive.

---

**Item 28:**

**Item stem:** Researchers at Renaissance Learning Inc. wanted to answer the following research question: Did the mathematics performance of Grade 2-8 students decline during the pandemic? A sample of students from across the U.S. took the Star assessments in fall 2019 (Grades 1-7) and fall 2020 (Grade 2-8). The scores were analyzed to determine which students were low performing and should either receive an educational intervention or be watched more closely. 42% of the students were considered low performing during the pandemic, as compared to 34% before the pandemic. Is there a need to conduct a hypothesis test to determine whether mathematics scores declined during the pandemic, or could we just use the sample statistic (42%) as evidence of such a decline?

Source: Renaissance Learning.

- We do not need to conduct a hypothesis test because 42% is much larger than 34%.
- We should conduct a hypothesis test because a hypothesis test is always appropriate.
- We should conduct a hypothesis test to determine if the sample statistic was unlikely to occur by chance.

---

**Items 29 and 30 refer to the following situation:**

The RAND Corporation regularly surveys people in the U.S. In their 2019 and 2020 surveys of the same 1,520 adults, one question asked was “In the past month (30 days), on how many days did you drink at least one full drink of alcohol?” One research question that the surveyors had was “Is there a difference between before COVID-19 pandemic and during COVID-19 pandemic behavior with regards to the average number of alcoholic drinks consumed?”

Source: Research article.

**Item 29:**

**Item stem:** Which of the following is a statement of the null hypothesis for a statistical test designed to answer the research question?

- There is no difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the number of alcoholic drinks consumed.
- There is a difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the number of alcoholic drinks consumed.
- There is no difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the average number of alcoholic drinks consumed.
- There is a difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the average number of alcoholic drinks consumed.

**Item 30:**

**Item stem:** Which of the following is a statement of the alternative hypothesis for a statistical test designed to answer the research question?

- There is no difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the number of alcoholic drinks consumed.
- There is a difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the number of alcoholic drinks consumed.
- There is no difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the average number of alcoholic drinks consumed.
- There is a difference between before COVID-19 pandemic and during COVID-19 pandemic behavior in terms of the average number of alcoholic drinks consumed.

---

**Item 31:**

**Item stem:** A scientist is designing a research study. They are hoping to show that the results of an experiment are statistically significant. What type of  $p$ -value would they want to obtain?

- A large  $p$ -value.
- A small  $p$ -value.
- The magnitude of a  $p$ -value has no impact on statistical significance.

---

**Item 32:**

**Item stem:** A clinical trial was conducted to determine if women who have regular mammograms to screen for breast cancer would decrease breast cancer mortality. The null hypothesis is women who have regular mammograms have the same breast cancer mortality rate as women who do not have regular mammograms. The alternative hypothesis is women who have regular mammograms have a lower breast cancer mortality rate than women who do not have regular mammograms. A hypothesis test was conducted and the results were not statistically significant. Does that mean that the null hypothesis is true, that women who have regular mammograms have the same breast cancer mortality rate as women who do not have regular mammograms?

- Yes. It means you cannot conclude that the alternative hypothesis is true, so the null hypothesis must be true.
- No. It means you cannot conclude that the null hypothesis is true, so the alternative hypothesis must be true.
- No. It means that there is not enough evidence to conclude that the null hypothesis is false.
- No. It means that there is not enough evidence to conclude that the alternative hypothesis is false.

---

**Item 33:**

**Item stem:** Dogs have a very strong sense of smell and have been trained to sniff various objects to pick up different scents. A pilot experiment was conducted with dogs in Germany who were trained to smell COVID-19 in saliva samples. In the test, one dog was presented with 115 saliva samples; 21 from COVID-19 patients and 94 from healthy people. The dog indicated which saliva samples were from the COVID-19 patients. Out of the 21 COVID-19 positive samples, the dog correctly identified 20 of them. A hypothesis test was conducted to see if this result could have happened by chance alone. The alternative hypothesis is that the dog correctly identifies COVID-19 more than half the times. The p-value is less than .001. Assuming it was a well-designed study, use a significance level of .05 to make a decision.

Source: Research article.

- Reject the null hypothesis and conclude that the dog correctly identifies COVID-19 more than half of the time.
- There is enough statistical evidence to prove that the dog correctly identifies COVID-19 more than half of the time.
- Do not reject the null hypothesis and conclude there is no evidence that the dog correctly identifies COVID-19 more than half of the time.

---

**Item 34:**

**Item stem:** Scientists are studying the relationship between the Oxford - AstraZeneca COVID-19 vaccine and a PCR test result positive for COVID-19. What type of study should they have conducted in order to establish that two doses of the vaccine cause a reduction in the chance of a positive PCR test?

Source: Study details.

- Observational study
- Randomized experiment
- Survey

---

**Item 35:**

**Item stem:** Gallup conducted a survey to measure people's well-being during the pandemic. They collected information regarding various life experiences and demographic

information. A random sample of adults from Poland were selected and 1,010 of them responded to the survey.

Which of the following does **NOT** affect Gallup's ability to generalize the survey results to the entire global adult population?

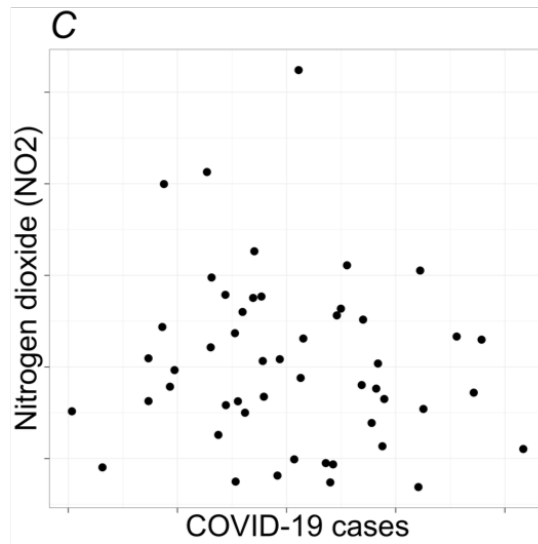
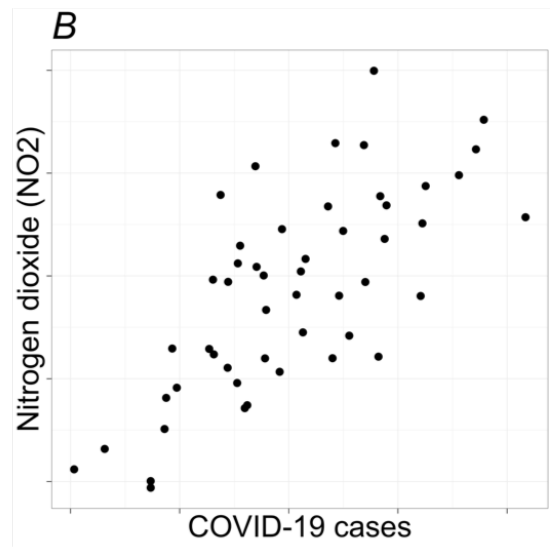
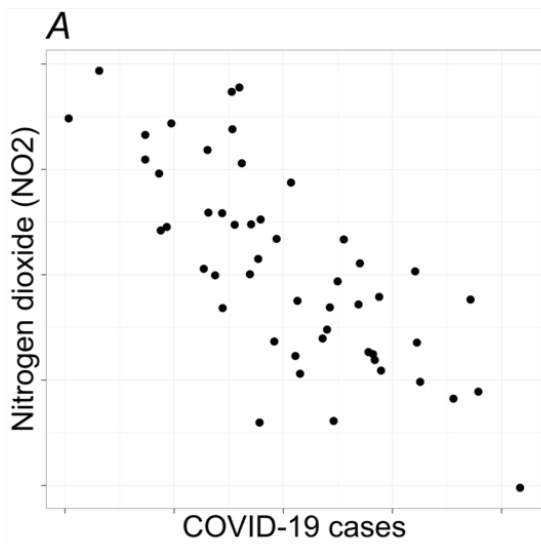
Source: World Happiness Report 2021.

- Although the total number of adults in the world is much higher, only 1,010 were surveyed.
- The survey was only given to Polish adults.
- Even though many more Polish adults were contacted, only 1,010 responded.
- All of the above present a problem for generalizing the results to people all over the world.

---

**Item 36:**

**Item stem:** Researchers studied the relationship between COVID-19 cases and air pollution in California, USA. They found that as cases increased, pollution measured in terms of Nitrogen dioxide (NO<sub>2</sub>) tended to reduce. Which of the following graphs illustrates this point?



Source: Research article.

- Graph A
- Graph B
- Graph C

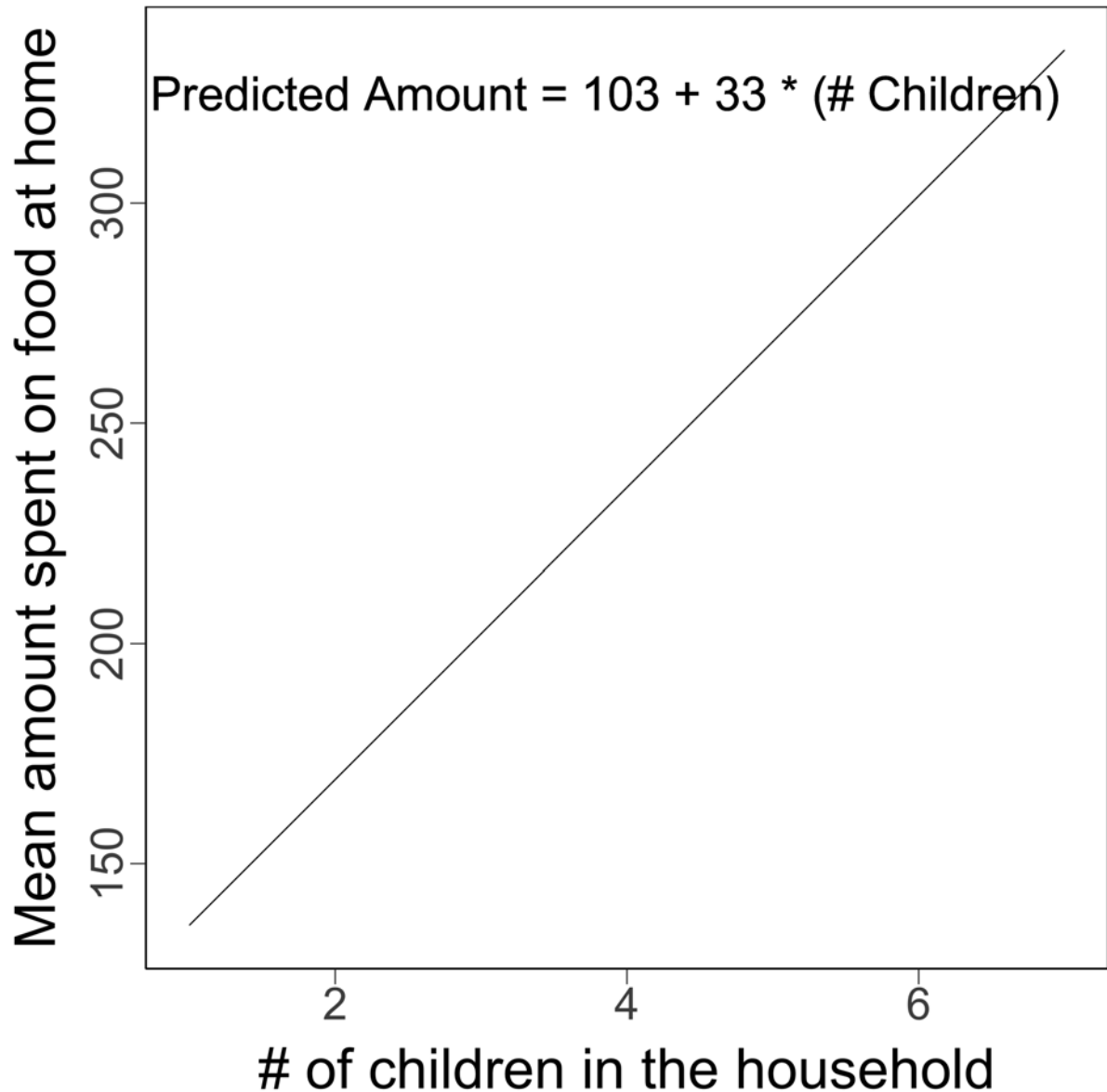
---

**Item 37:**

**Item stem:** The U.S. Census Bureau launched the Household Pulse Survey to study the impact of the COVID-19 pandemic on households across the country. In the most



recent round, they gathered data from over 78,000 households. Using the data on Number of Children in the Household and Mean Amount Spent on Food Prepared and Eaten at Home (in U.S. dollars) during the previous week, they found a linear relationship and produced the following regression equation and plot of the regression equation:



Suppose you are asked to use regression to predict the mean amount spent on food at home for a household with 5 children. Which of the following methods can be used to provide an estimate?

Source: Census Bureau.

- Locate the point on the line that corresponds to 5 children in the household and read off the corresponding value on the y axis.
- Substitute 5 for (# Children) in the equation and solve for "Predicted Amount".
- Both of these methods are correct.
- Neither of these methods is correct.

# Appendix B |

## Additional Results for Chapter 2

### B.1 Respondent Demographics

#### B.1.1 Univariate summaries

Gender	Frequency
Woman	0.605
Man*	0.387
Transgender	0.003
Prefer not to disclose	0.003
Prefer to self-specify	0.002

Table B.1: Gender identification: n = 1253

Whether an international student	Frequency
No*	0.931
Yes	0.069

Table B.2: International student: n = 1253

Class standing	Frequency
First Year (e.g. Freshman)*	0.638
Second Year (e.g. Sophomore)	0.228
Third Year (e.g. Junior)	0.097
Fourth Year or Higher (e.g. Senior)	0.037

Table B.3: Class standing: n = 1253

Prior statistics training	Frequency
No*	0.696
Yes	0.304

Table B.4: Prior statistics training: n = 1253

Expected course grade	Frequency
A*	0.338
B	0.410
C	0.216
D	0.032
F	0.004

Table B.5: Expected course grade: n = 1253

Highest education level of a parent/guardian	Frequency
Less than high school*	0.011
High school graduate	0.080
Some college, no degree	0.080
Associates Degree	0.041
Bachelor's Degree	0.391

Highest education level of a parent/guardian	Frequency
Some graduate school	0.025
Master's Degree	0.260
Professional Degree	0.036
Doctorate Degree	0.076

Table B.6: Highest education of parent/guardian: n = 1253

## B.2 Assessment Response Summaries

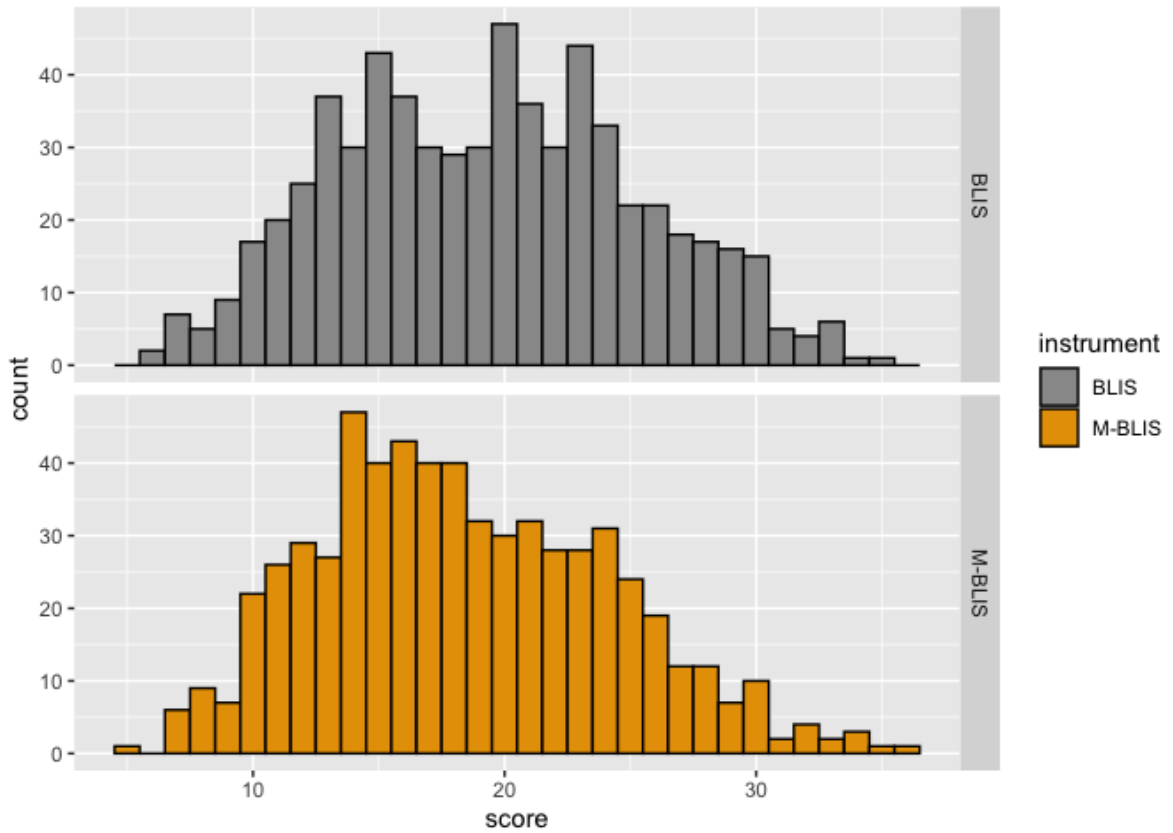


Figure B.1. Comparison of total score (out of 37) - separate panels

Item	BLIS				M-BLIS			
	A	B	C	D	A	B	C	D
1	16.6	8.8	74.6*	NA	19.3	7.5	73.2*	NA
2	7.4	15.8	44*	32.8	8.6	18	50.7*	22.6
3	53.4*	20.5	7.2	18.8	52.5*	8.6	13	25.9
4	15.7	83.5*	0.8	NA	13	86.2*	0.8	NA
5	81.3*	16.5	2.2	NA	84.7*	13.8	1.5	NA
6	3.4	10.8	73.5*	12.2	4.7	15.8	70.7*	8.8
7	32.6	15.7	16.1	35.6*	21	17.1	20.8	41.1*
8	39.8	30.7	29.5*	NA	23.7	43.4	32.8*	NA
9	65.4*	30.7	3.9	NA	34*	56.4	9.6	NA
10	14.7	9.1	19.9	56.3*	21.6	21.5	17.7	39.2*
11	19.1	38.9	42*	NA	18.5	44.4	37.1*	NA
12	13.9	22.7	5	58.3*	11.9	31.1	8.3	48.8*
13*	37.6*	40.3	8.5	13.6	37.6*	35	14.1	13.3
14	5.8	42.8*	50.3	1.1	2.8	69.1	24.6*	3.6
15	5.6	8.9	21.6	63.8*	6	18.2	27.3	48.5*
16*	24.6*	19.9	36.1	19.4	27.8*	42.4	29.8	NA
17*	25.2	28.7	46.1*	NA	23.7	29.4	46.8*	NA
18	35.3	45.9*	18.8	NA	29.9	45.4*	24.7	NA
19	9.7	29.8	19.7	40.8*	11.1	27.6	22.9	38.4*
20	20.7	29	37.9*	12.4	17.9	34.8	34.3*	13
21	16.5*	8.8	39.3	35.4	16.3*	14.1	36.7	32.8
22	29.6	58.5*	11.9	NA	26.8	61*	12.2	NA
23*	12.2	29.8	43.4*	14.6	10.4	31.1	43.9*	14.6
24*	57.2*	25.4	17.4	NA	60*	22.4	17.6	NA
25	55.5*	25.9	18.7	NA	61.6*	28	10.4	NA
26	42.2*	42.5	15.4	NA	42*	42.3	15.8	NA
27	31.7	38.6*	18.8	11	27.5	45*	14.6	12.8
28	20.5	26.8	52.7*	NA	17.2	22.6	60.2*	NA
29	18	17.4	52*	12.5	13.3	20.2	48.9*	17.6
30	10	19.9	21.8	48.3*	9.6	22.4	22.4	45.5*
31	8.2	86.4*	5.5	NA	9.9	83.9*	6.2	NA
32*	19.7	20.8	48*	11.4	20	23.6	43.6*	12.8
33	64.4*	22.1	13.5	NA	62*	25.5	12.5	NA

Item	BLIS				M-BLIS			
	A	B	C	D	A	B	C	D
34	25.4	70.4*	4.2	NA	31.2	65.2*	3.6	NA
35	23.4*	11.4	13	52.2	21.6*	16.9	15.4	46
36	79*	16.6	4.4	NA	68.6*	23.7	7.6	NA
37	16.6	20.7	57.8*	4.9	19.8	20.8	54.3*	5

Table B.7: Selected-response table

### B.3 Reliability and Validity Evidence

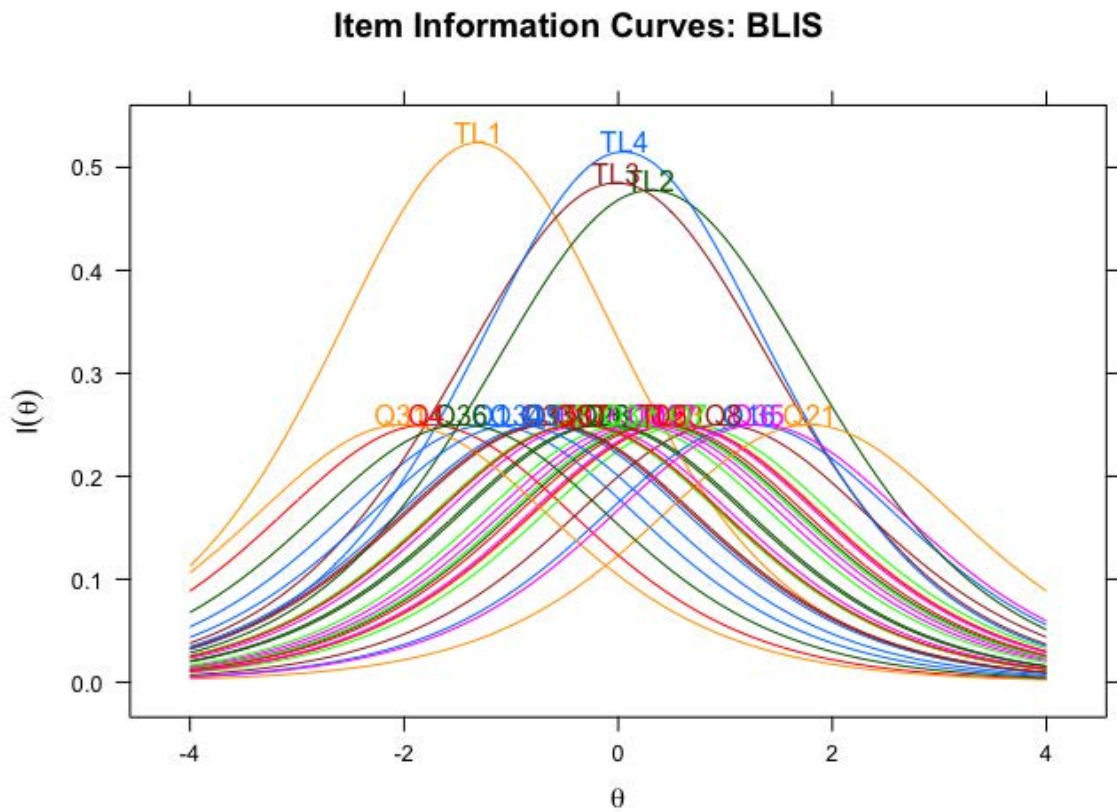


Figure B.2. Item Information Curves - BLIS

### Item Information Curves: M-BLIS

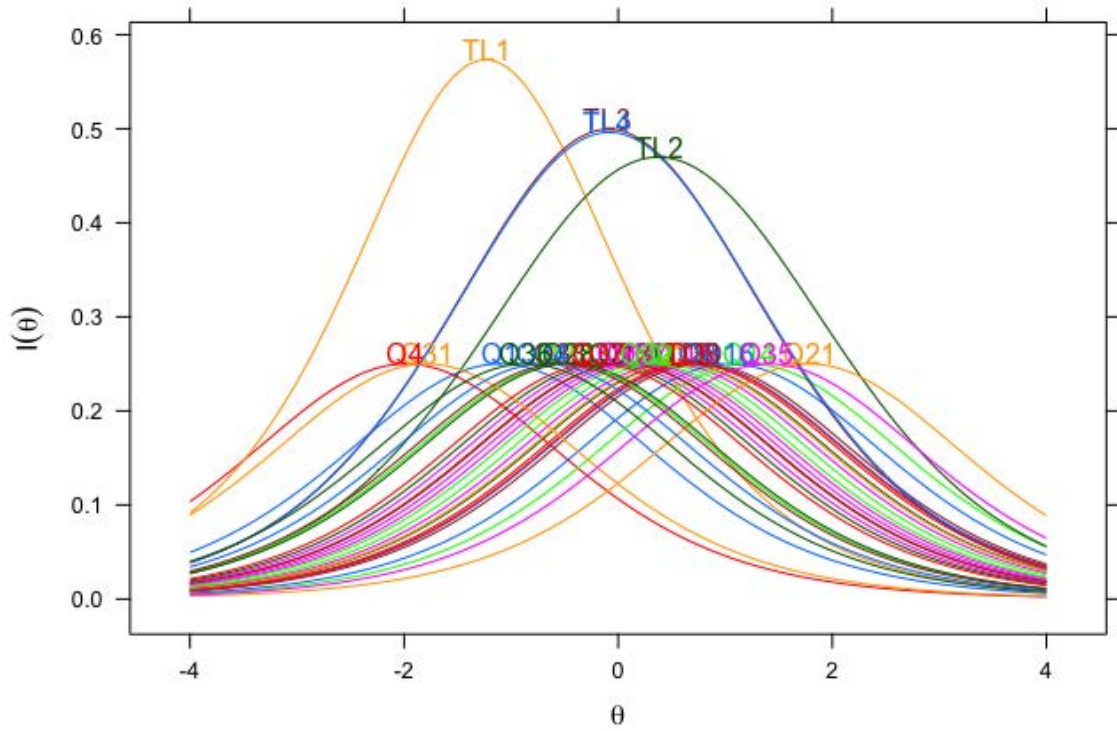
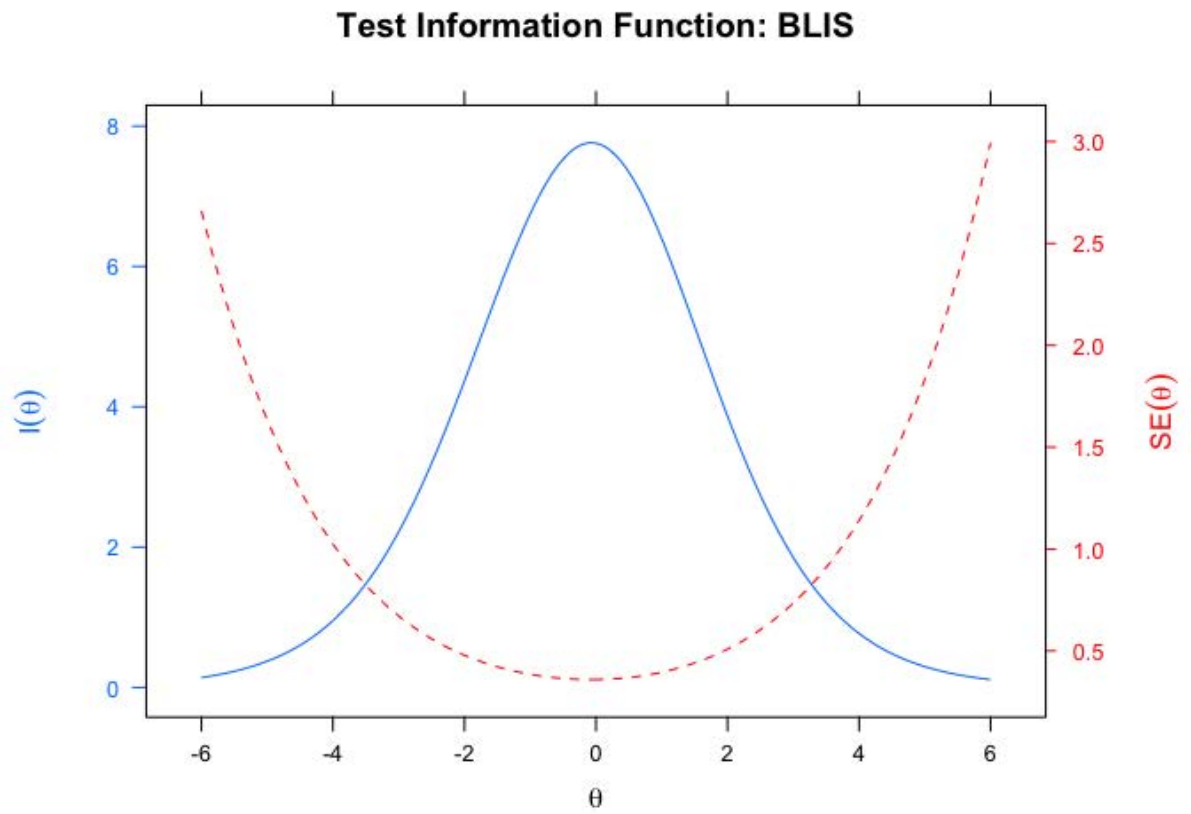
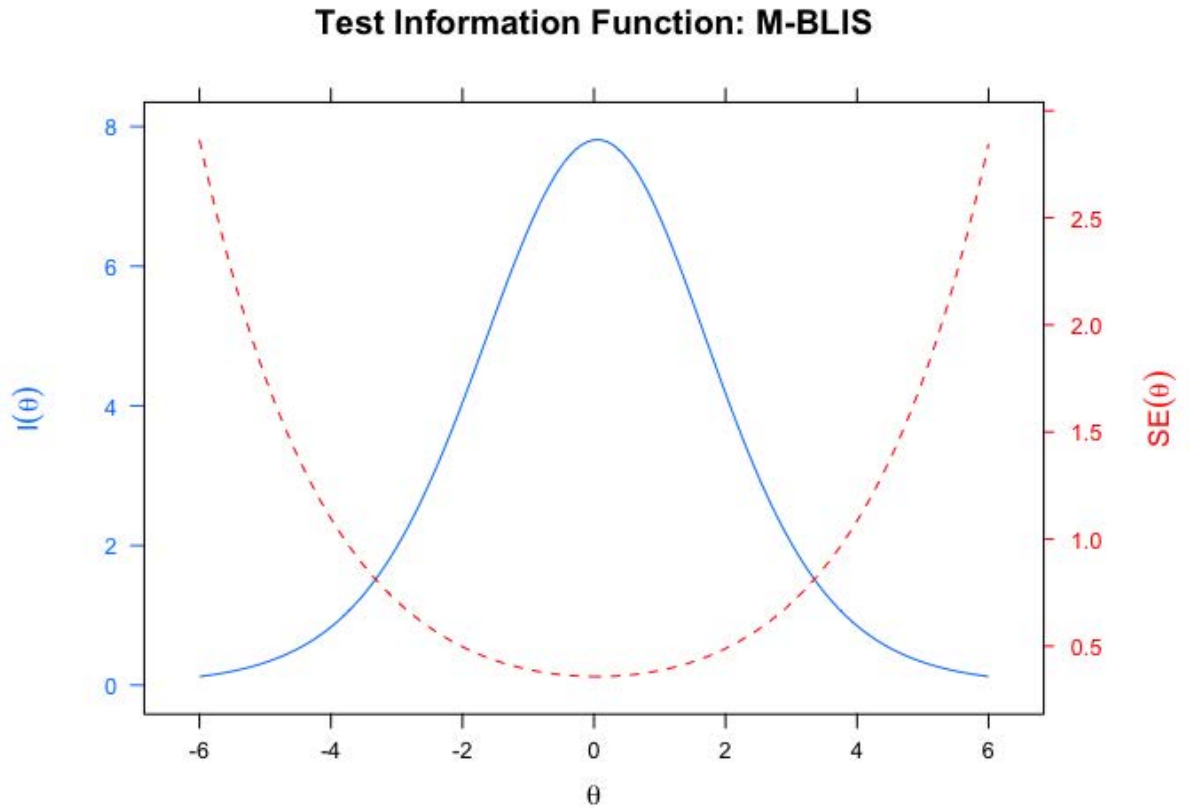


Figure B.3. Item Information Curves - M-BLIS





**Figure B.4.** Test Information Function and Standard Error - BLIS



**Figure B.5.** Test Information Function and Standard Error - M-BLIS

Item	BLIS	M-BLIS
Q1	-1.1927156	-1.1021233
Q2	0.2631418	-0.0375388
Q3	-0.1571817	-0.1164450
Q4	-1.7811054	-1.9824458
TL1	-1.3170586	-1.2386935
Q7	0.6567262	0.3897469
Q8	0.9650827	0.7822983
Q9	-0.7077156	0.7261573
Q10	-0.2838843	0.4792098
Q11	0.3556544	0.5778552
Q12	-0.3763797	0.0485164
Q13	0.5593310	0.5549103
Q14	0.3199521	1.2271370

Item	BLIS	M-BLIS
Q15	-0.6321048	0.0628702
Q16	1.2357476	1.0437791
Q17	0.1714903	0.1347889
TL2	0.3055317	0.3746852
Q20	0.5445291	0.7102782
Q21	1.7825288	1.7836447
Q22	-0.3835169	-0.4960496
TL3	-0.0174983	-0.0913839
TL4	0.0476351	-0.0845298
Q27	0.5150569	0.2142002
Q28	-0.1221481	-0.4587553
TL5	0.4059238	0.6708731
Q31	-2.0158467	-1.7937093
Q32	0.0873939	0.2796047
Q33	-0.6621846	-0.5411665
Q34	-0.9608257	-0.6947486
Q35	1.3108151	1.4059160
Q36	-1.4599263	-0.8628098
Q37	-0.3549298	-0.1956216

Table B.8: Difficulty estimates based on PC model

### Item Characteristic Curves: BLIS

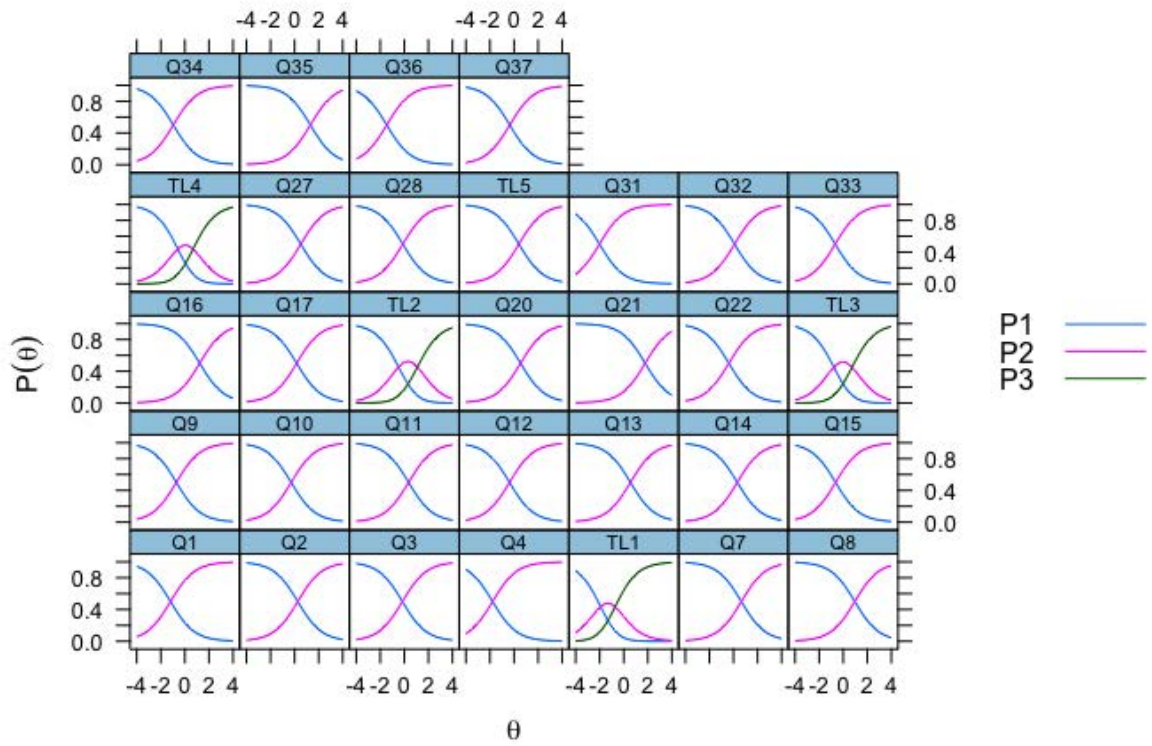


Figure B.6. Item Characteristic Curves - BLIS

### Item Characteristic Curves: M-BLIS

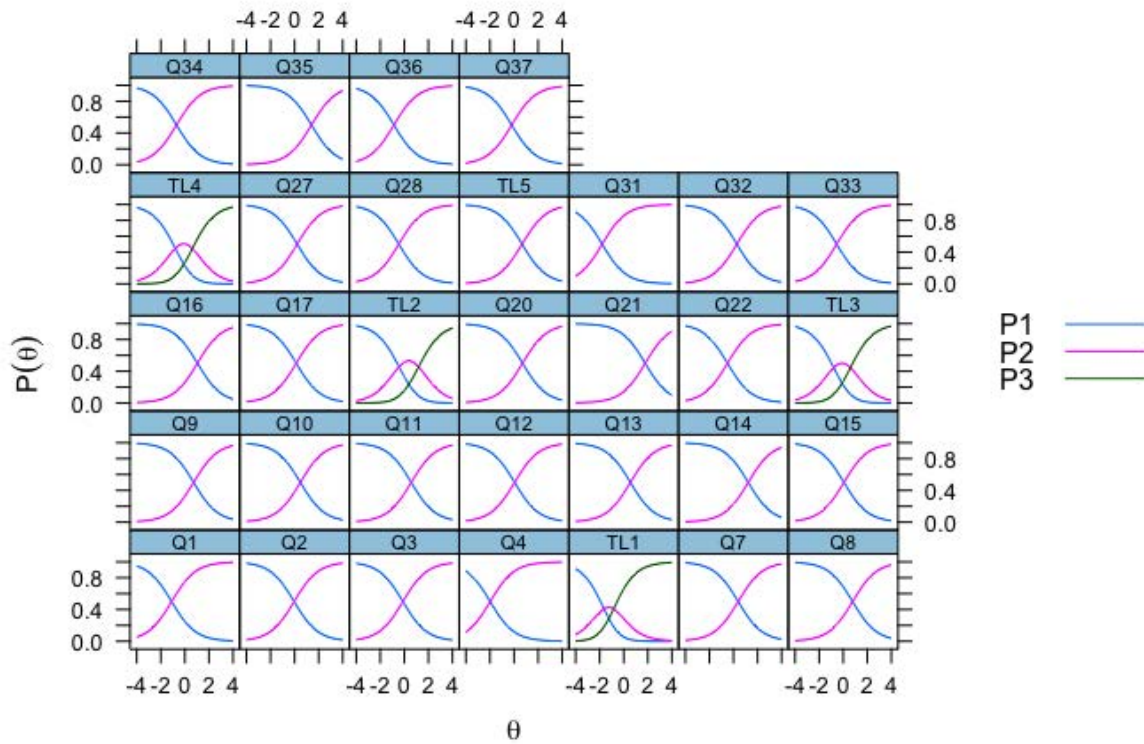


Figure B.7. Item Characteristic Curves - M-BLIS

## B.4 Regression Results

	Estimate	Std. Error	t-value	p-value
(Intercept)	25.43	2.31	11.03	0.0000
Instrument - M-BLIS	-1.21	0.33	-3.64	0.0003
International - Yes	-1.27	0.75	-1.68	0.0937
Grade - B	-4.57	0.38	-11.95	0.0000
Grade - C	-6.88	0.49	-13.99	0.0000
Grade - D	-8.77	0.99	-8.87	0.0000
Grade - F	-0.28	3.72	-0.07	0.9407
priorSTAT - Yes	0.56	0.37	1.53	0.1255
Class - Second Year (e.g. Sophomore)	-0.31	0.41	-0.76	0.4479
Class - Third Year (e.g. Junior)	-0.25	0.63	-0.39	0.6980
Class - Fourth Year or Higher (e.g. Senior)	1.77	1.02	1.74	0.0823

	Estimate	Std. Error	t-value	p-value
Gender - Woman	-0.50	0.35	-1.43	0.1524
Gender - Transgender	-2.37	3.74	-0.64	0.5256
Gender - Prefer not to disclose	2.84	3.67	0.77	0.4391
Gender - Prefer to self-specify	4.18	5.17	0.81	0.4182
Highest parent ed - High school graduate	-2.72	1.96	-1.39	0.1662
Highest parent ed - College, no degree	-1.96	1.95	-1.01	0.3144
Highest parent ed - Associate's	-1.75	2.06	-0.85	0.3966
Highest parent ed - Bachelor's	-1.51	1.89	-0.80	0.4236
Highest parent ed - Some graduate school	0.14	2.14	0.07	0.9470
Highest parent ed - Master's	-0.70	1.90	-0.37	0.7112
Highest parent ed - Professional degree	-1.26	2.07	-0.61	0.5415
Highest parent ed - Doctorate	-0.41	1.95	-0.21	0.8338
COVID engagement - Maybe	-0.38	0.91	-0.42	0.6726
COVID engagement - Yes	0.03	0.68	0.04	0.9681
COVID interest - Maybe	1.22	0.55	2.21	0.0274
COVID interest - Yes	0.39	0.49	0.80	0.4245
COVID relevance - Maybe	-3.59	1.48	-2.42	0.0155
COVID relevance - Yes	-0.92	1.15	-0.80	0.4260
Familiarity with topic	-0.01	0.01	-1.39	0.1655
Interest in topic	0.01	0.01	1.51	0.1324
Context - made question easier	-0.17	0.35	-0.48	0.6341
Context - made question harder	-0.29	0.93	-0.32	0.7520

Table B.9: Results from the full regression model

	Estimate	Std. Error	t value	p-value
(Intercept)	28.89	3.83	7.53	0.0000
Instrument - M-BLIS	-5.80	4.81	-1.20	0.2286
International - Yes	0.0788	1.1140	0.0708	0.8752
Grade - B	-4.5921	0.5257	-8.7348	0.0000
Grade - C	-6.7615	0.6682	-10.1184	0.0000
Grade - D	-8.5514	1.2567	-6.8048	0.0000
Grade - F	0.4755	3.7834	0.1257	0.9096

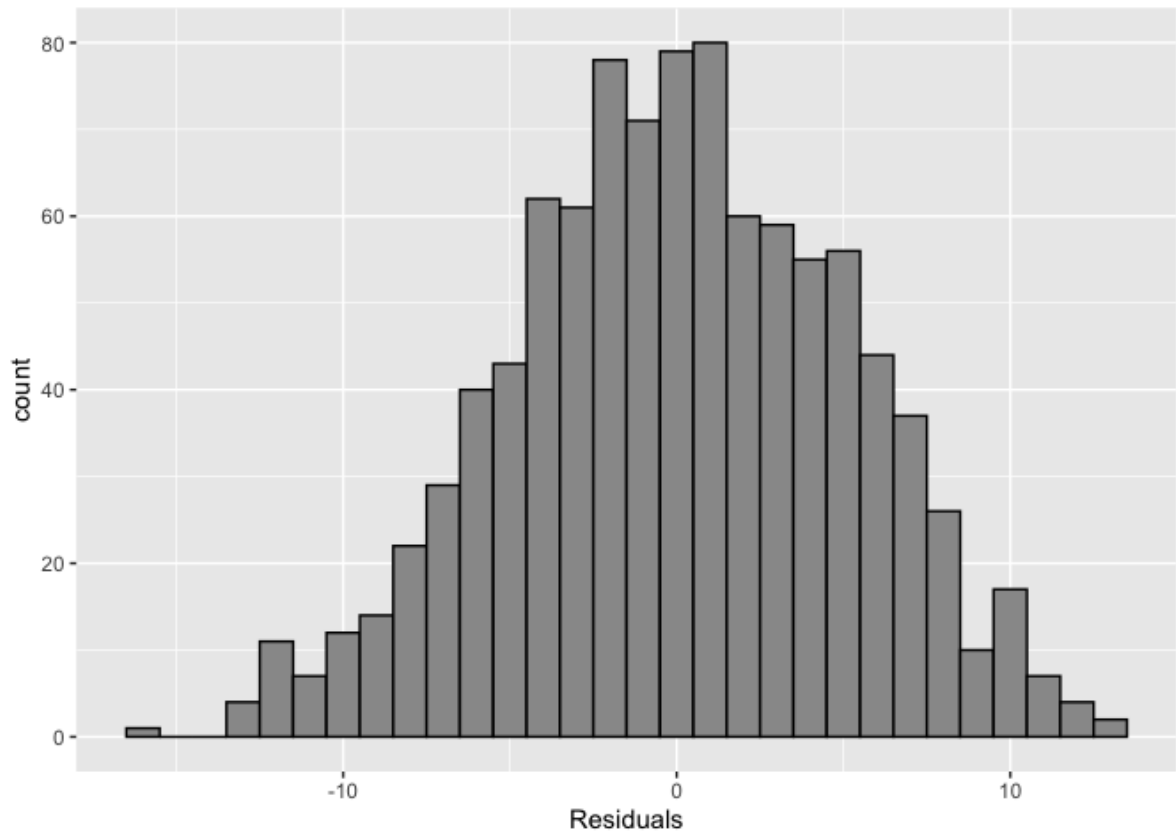
	Estimate	Std. Error	t value	p-value
priorSTAT - Yes	1.1996	0.4903	2.4465	0.0172
Class - Second Year (e.g. Sophomore)	0.2548	0.5755	0.4428	0.8956
Class - Third Year (e.g. Junior)	-0.2159	0.8732	-0.2472	0.6428
Class - Fourth Year or Higher (e.g. Senior)	1.2674	1.4629	0.8664	0.0870
Gender - Woman	-0.2991	0.4722	-0.6335	0.4736
Gender - Transgender	-1.4369	3.8185	-0.3763	0.7167
Gender - Prefer not to disclose	3.4160	3.6900	0.9257	0.3400
Gender - Prefer to self-specify	4.0562	5.1820	0.7827	0.4726
Highest parent ed - High school graduate	-3.2444	2.8007	-1.1584	0.0548
Highest parent ed - College, no degree	-3.8873	2.8068	-1.3850	0.0329
Highest parent ed - Associate's	-4.0512	2.9570	-1.3700	0.0376
Highest parent ed - Bachelor's	-3.0528	2.7370	-1.1154	0.0584
Highest parent ed - Some graduate school	0.3851	3.0209	0.1275	0.5557
Highest parent ed - Master's	-1.4848	2.7361	-0.5427	0.1755
Highest parent ed - Professional degree	-0.9369	2.9184	-0.3210	0.2416
Highest parent ed - Doctorate	-0.9873	2.8362	-0.3481	0.2580
COVID engagement - Maybe	0.1007	1.2342	0.0816	0.7519
COVID engagement - Yes	0.6973	0.9523	0.7322	0.2651
COVID interest - Maybe	0.6063	0.7670	0.7904	0.4399
COVID interest - Yes	0.2437	0.6822	0.3573	0.8167
COVID relevance - Maybe	-4.5675	2.1554	-2.1191	0.0688
COVID relevance - Yes	-2.6750	1.8918	-1.4140	0.1646
Familiarity with topic	0.0000	0.0136	-0.0010	0.8783
Interest in topic	0.0100	0.0129	0.7748	0.5249
Context - made question easier	0.7601	0.4796	1.5850	0.1034
Context - made question harder	-0.2331	1.3634	-0.1710	0.8917
M-BLIS * International - Yes	-2.1982	1.5346	-1.4324	0.1621
M-BLIS * Grade - B	0.1978	0.7571	0.2613	0.9379
M-BLIS * Grade - C	0.0312	0.9662	0.0323	0.9841
M-BLIS * Grade - D	-0.4991	1.9815	-0.2519	0.7459
M-BLIS * priorSTAT - Yes	-1.4494	0.7287	-1.9891	0.0687
M-BLIS * Class - Second Year	-1.1473	0.8154	-1.4071	0.3645
M-BLIS * Class - Third Year	-0.2476	1.2452	-0.1989	0.8984
M-BLIS * Class - Fourth Year or Higher	0.0364	1.9530	0.0186	0.4287

	Estimate	Std. Error	t value	p-value
M-BLIS * Gender - Woman	-0.3288	0.6920	-0.4751	0.5639
M-BLIS * Parent ed - High school graduate	1.4023	3.8061	0.3684	0.2728
M-BLIS * Parent ed - College, no degree	4.1724	3.7799	1.1038	0.0754
M-BLIS * Parent ed - Associate's	5.3618	4.0086	1.3376	0.0519
M-BLIS * Parent ed - Bachelor's	3.2610	3.6685	0.8889	0.1024
M-BLIS * Parent ed - Some graduate school	-1.3695	4.1521	-0.3298	0.7980
M-BLIS * Parent ed - Master's	1.8308	3.6731	0.4984	0.2220
M-BLIS * Parent ed - Professional degree	0.2210	4.0395	0.0547	0.4898
M-BLIS * Parent ed - Doctorate	1.6842	3.7887	0.4445	0.2900
M-BLIS * COVID engagement - Maybe	-1.4961	1.7527	-0.8536	0.4441
M-BLIS * COVID engagement - Yes	-1.2941	1.3267	-0.9755	0.1871
M-BLIS * COVID interest - Maybe	0.9790	1.0848	0.9025	0.2342
M-BLIS * COVID interest - Yes	-0.0937	0.9731	-0.0963	0.7556
M-BLIS * COVID relevance - Maybe	0.8075	3.0687	0.2631	0.9063
M-BLIS * COVID relevance - Yes	2.9060	2.3986	1.2115	0.2139
M-BLIS * Familiarity with topic	-0.0179	0.0201	-0.8892	0.4097
M-BLIS * Interest in topic	0.0073	0.0190	0.3855	0.7339
M-BLIS * Context - made question easier	-1.9502	0.6933	-2.8130	0.0065
M-BLIS * Context - made question harder	-0.5543	1.8714	-0.2962	0.5633

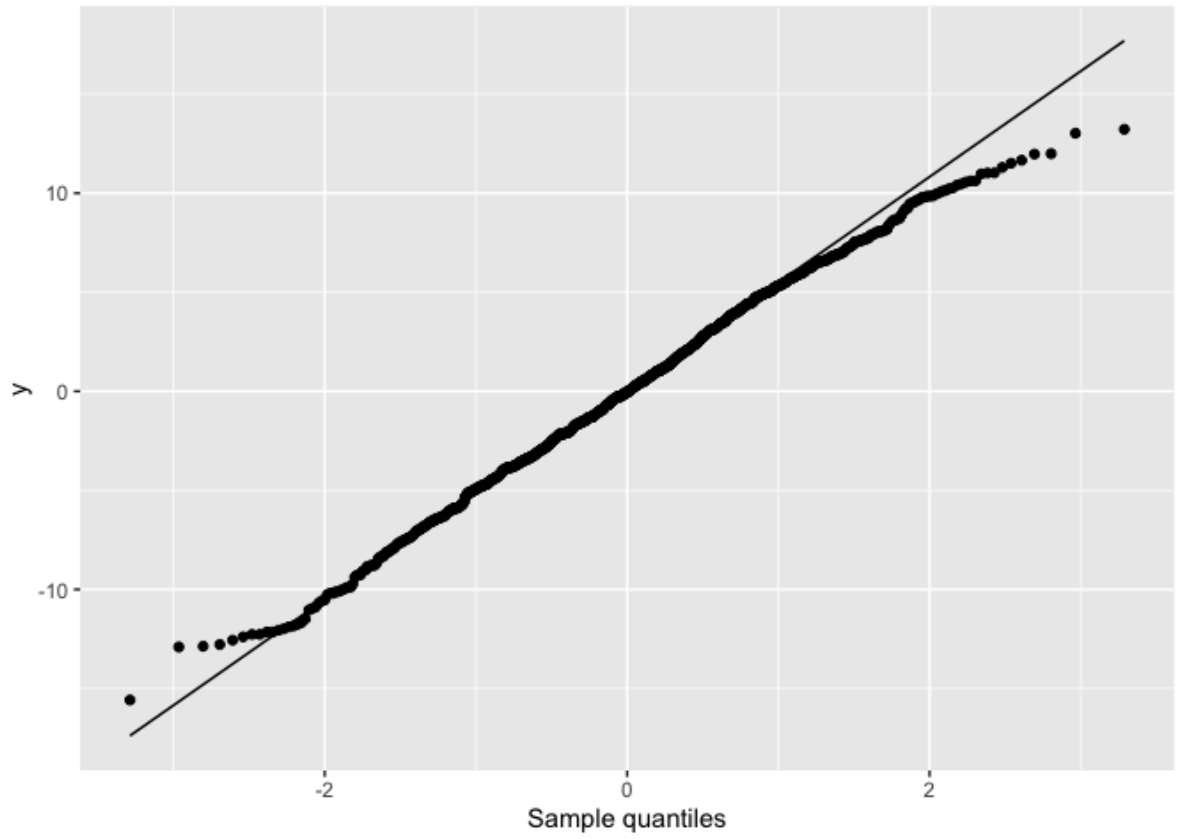
Table B.10: Results from the full regression model with interactions



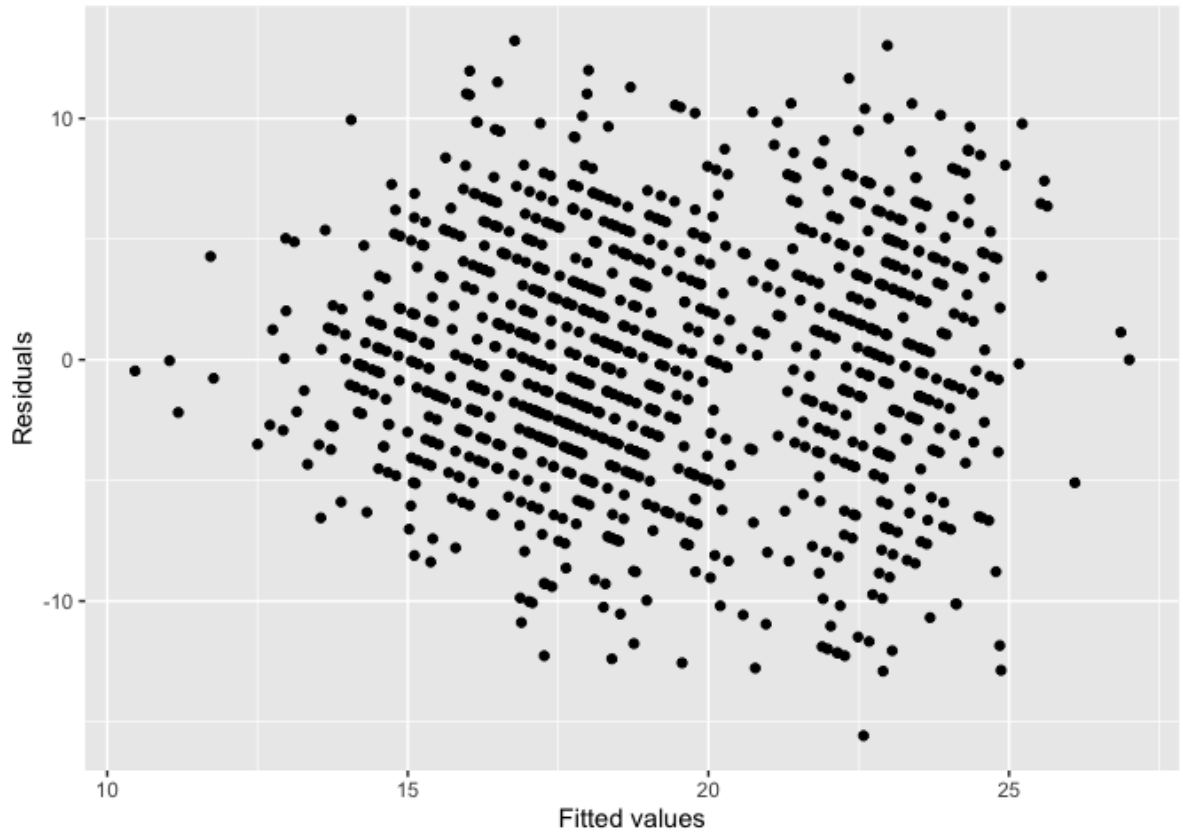
### B.4.1 Diagnostic plots for the additive model



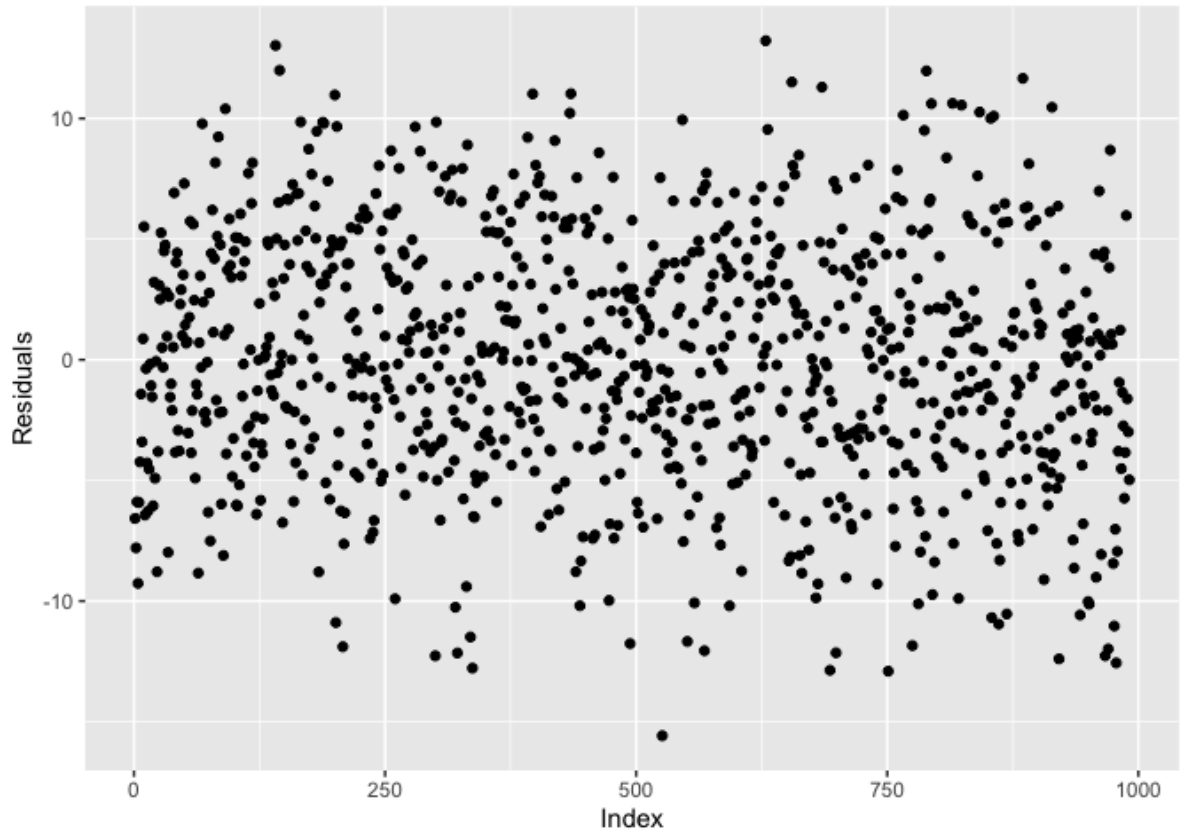
**Figure B.8.** Histogram of residuals - Full model in Equation 1



**Figure B.9.** Quartile-quartile plot of residuals - Full model in Equation 1

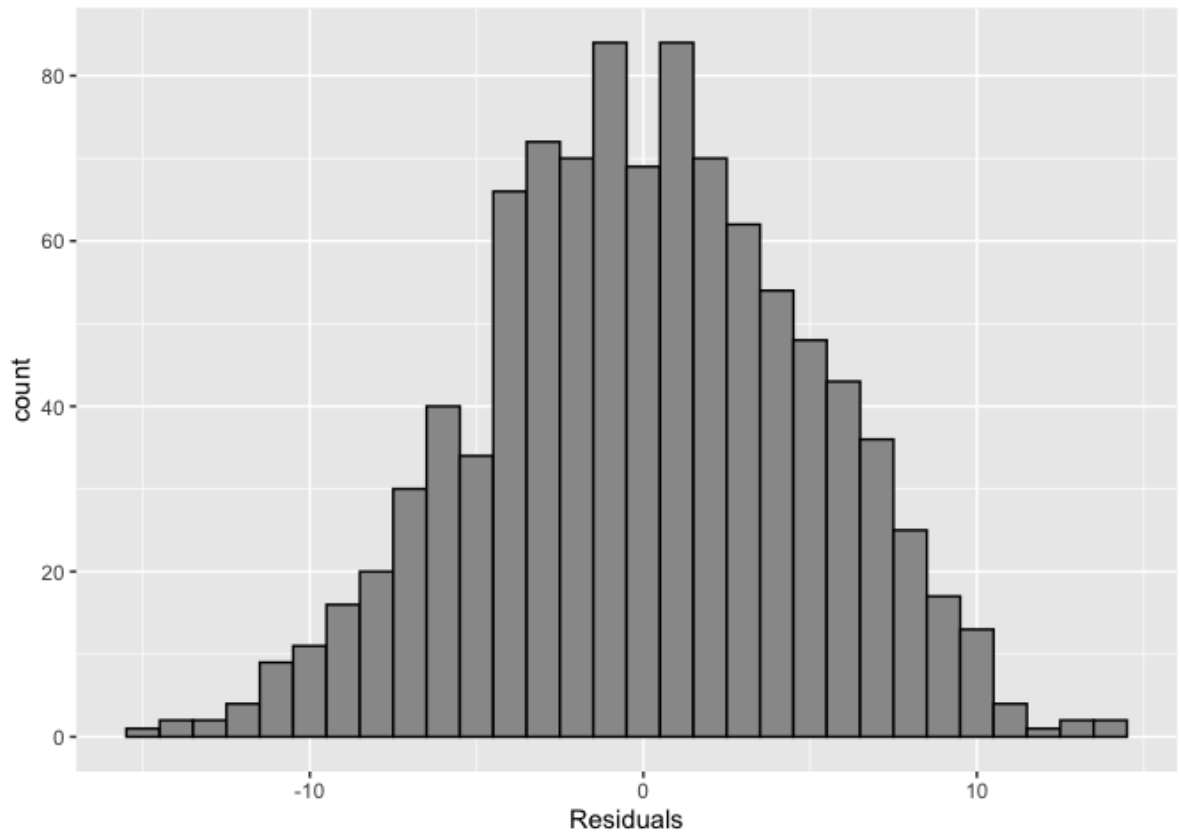


**Figure B.10.** Fitted values versus residuals plot - Full model in Equation 1

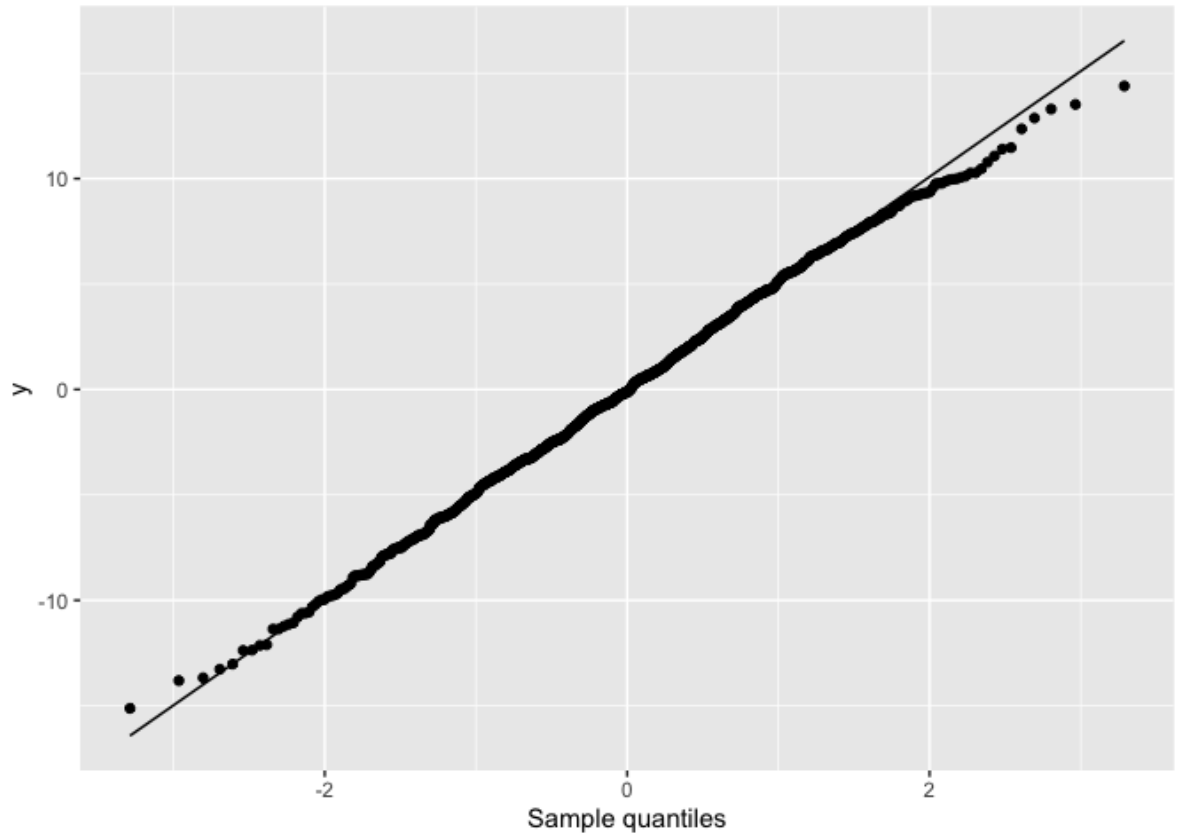


**Figure B.11.** Residuals plot - Full model in Equation 1

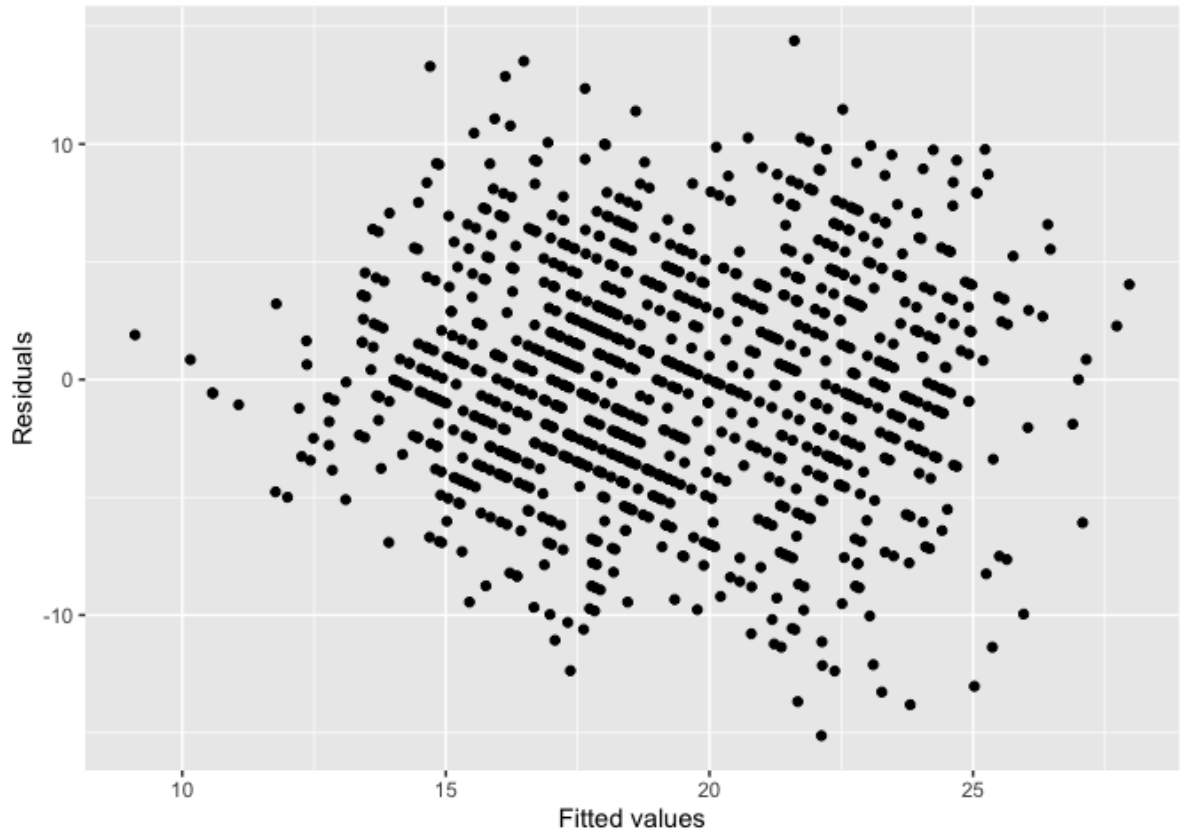
## B.4.2 Diagnostic plots for the interaction model



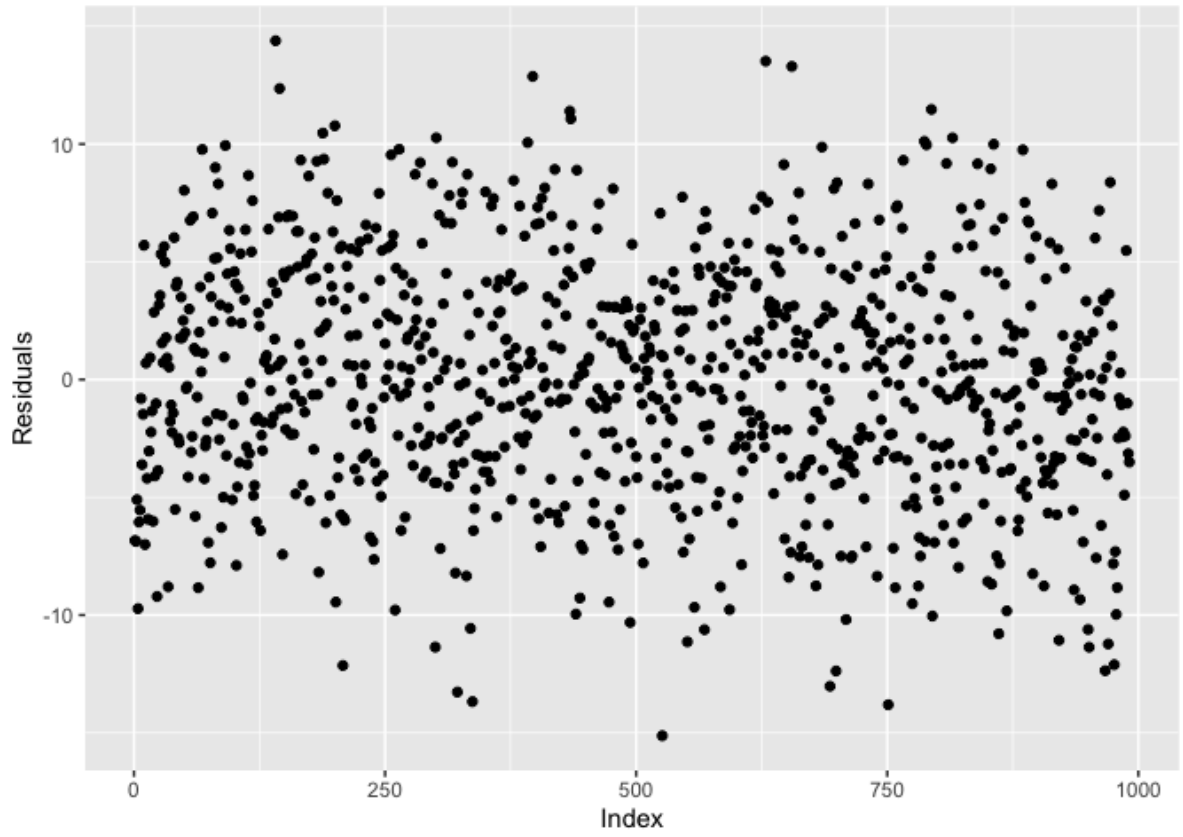
**Figure B.12.** Histogram of residuals - Full model plus interactions



**Figure B.13.** Quartile-quartile plot of residuals - Full model plus interactions



**Figure B.14.** Fitted values versus residuals plot - Full model plus interactions



**Figure B.15.** Residuals plot - Full model plus interactions



# Appendix C | Additional Results for Chapter 3

## C.1 Descriptive Summaries

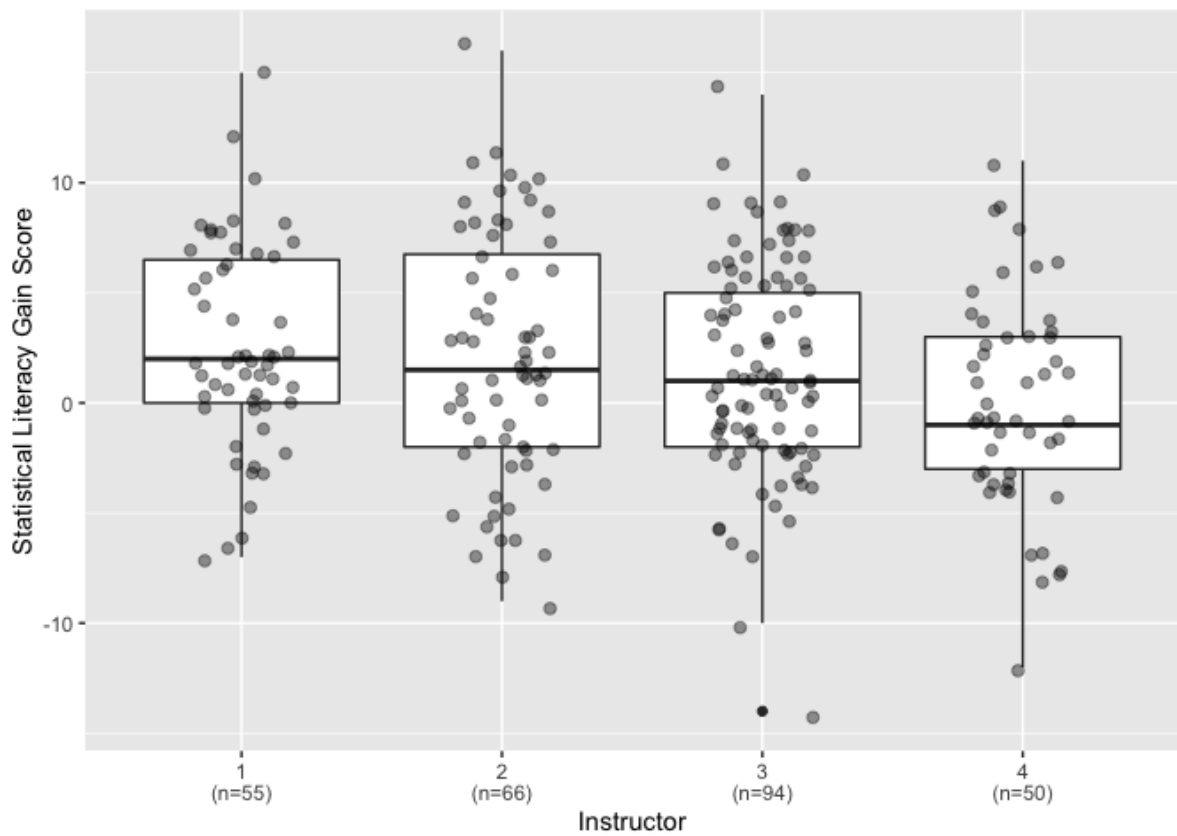


Figure C.1. Boxplot of gain score by instructor

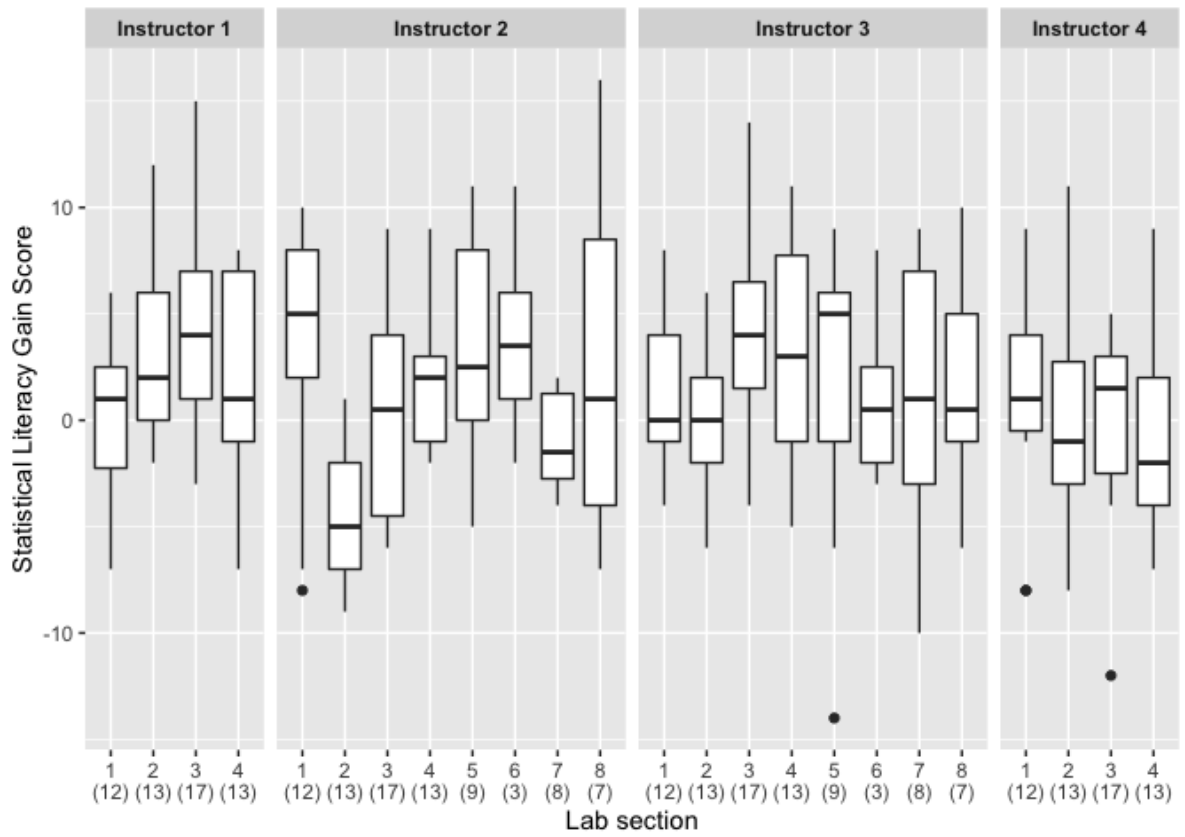
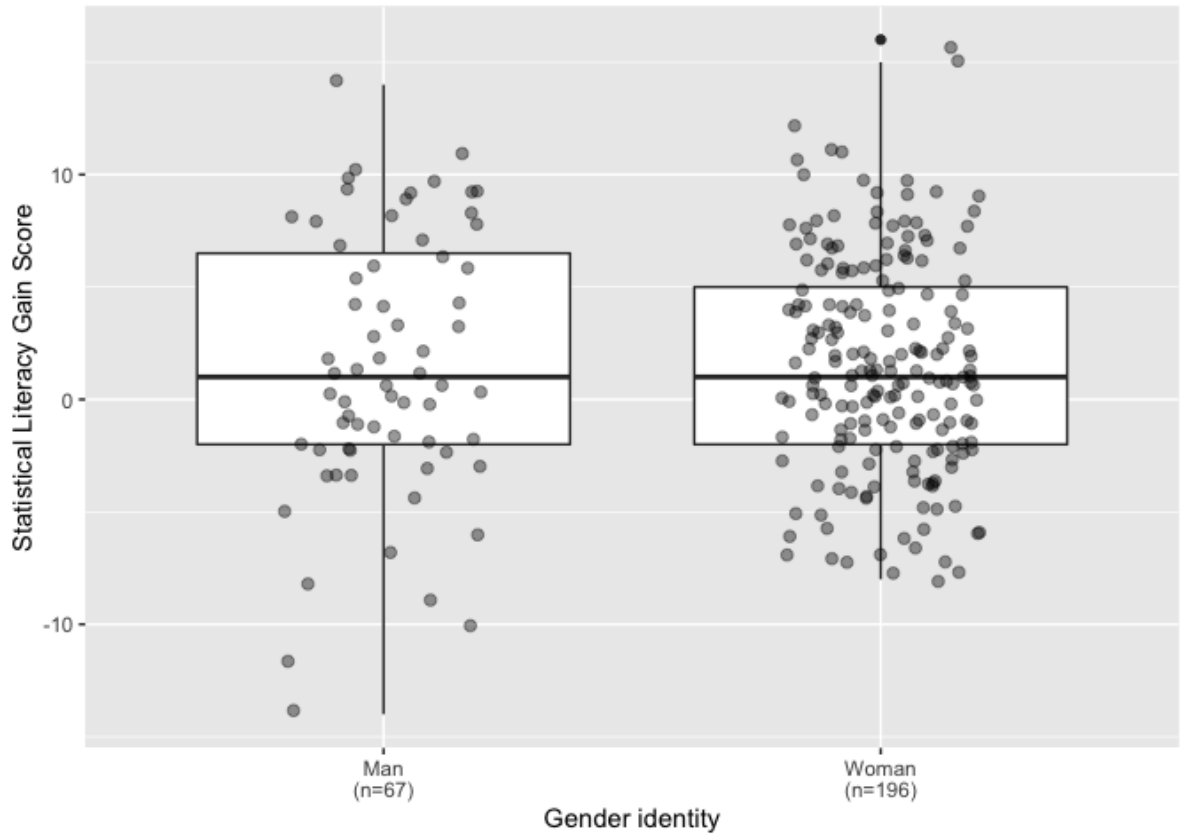
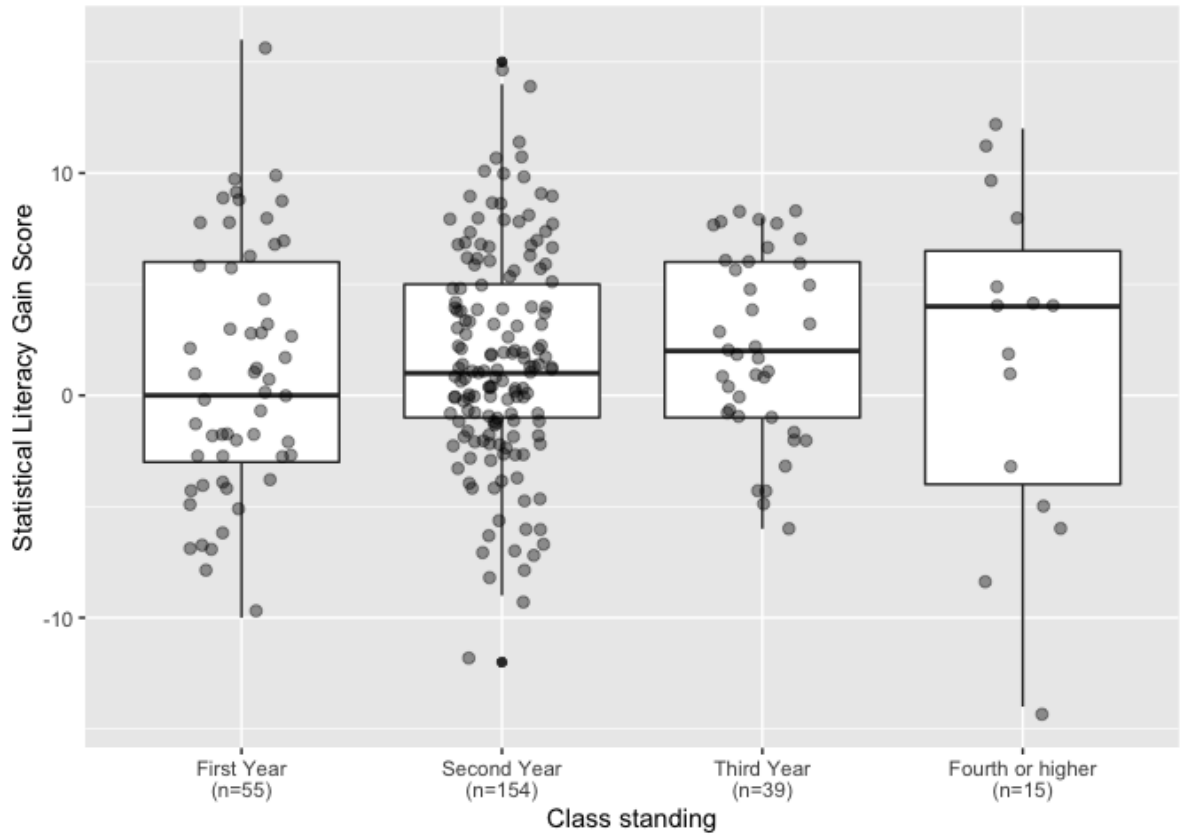


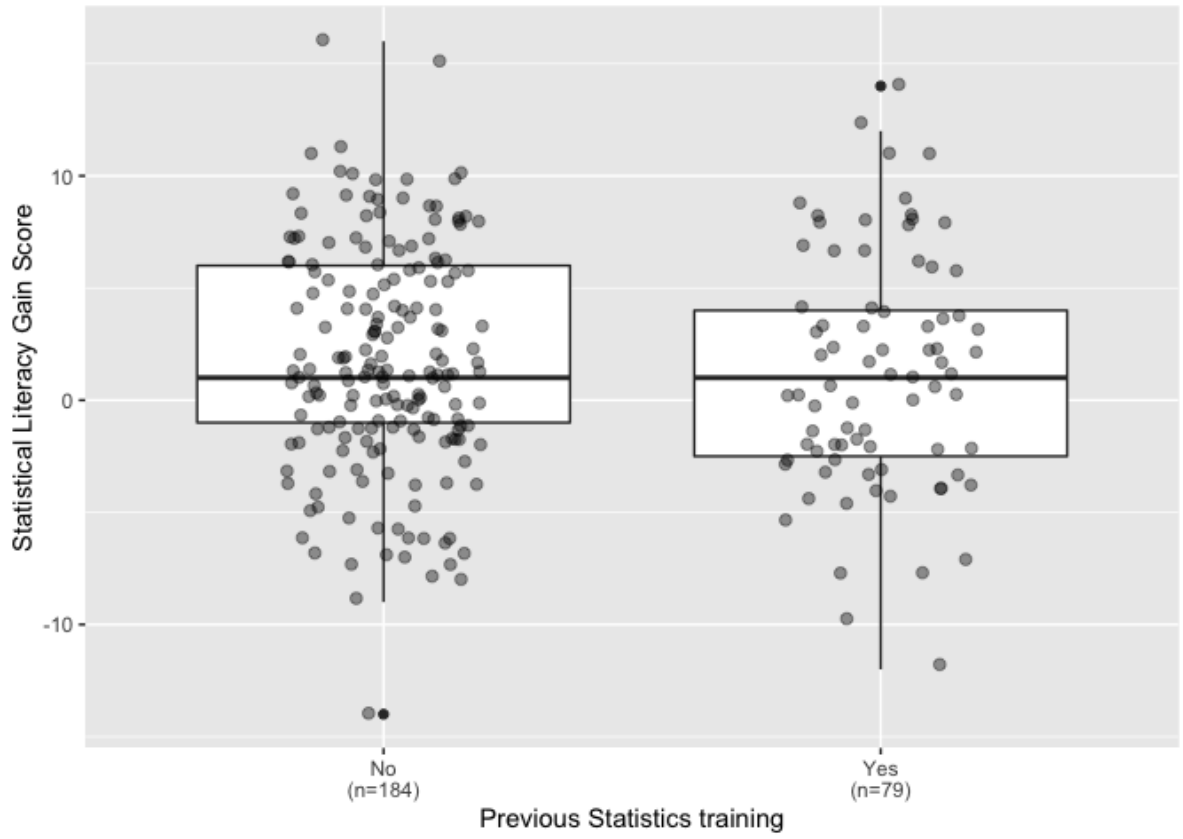
Figure C.2. Boxplot of gain score by lab sections



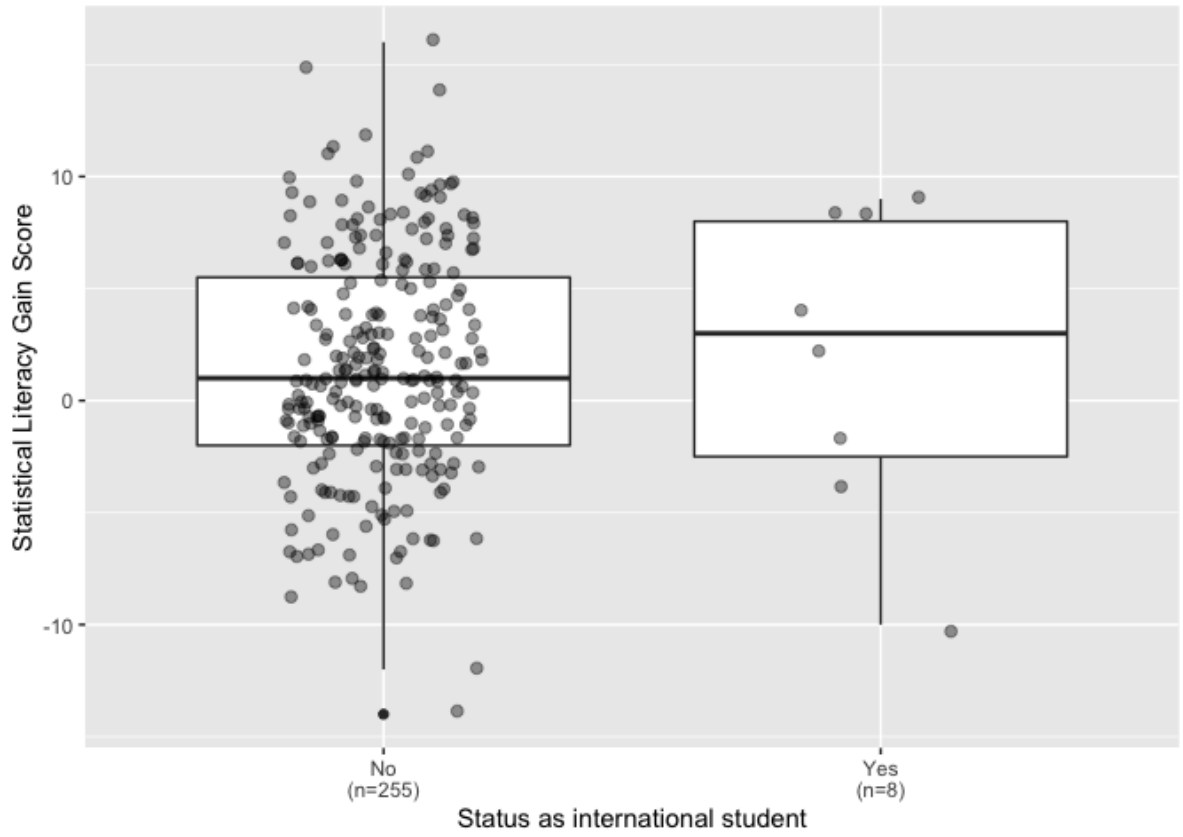
**Figure C.3.** Boxplot of gain score by gender



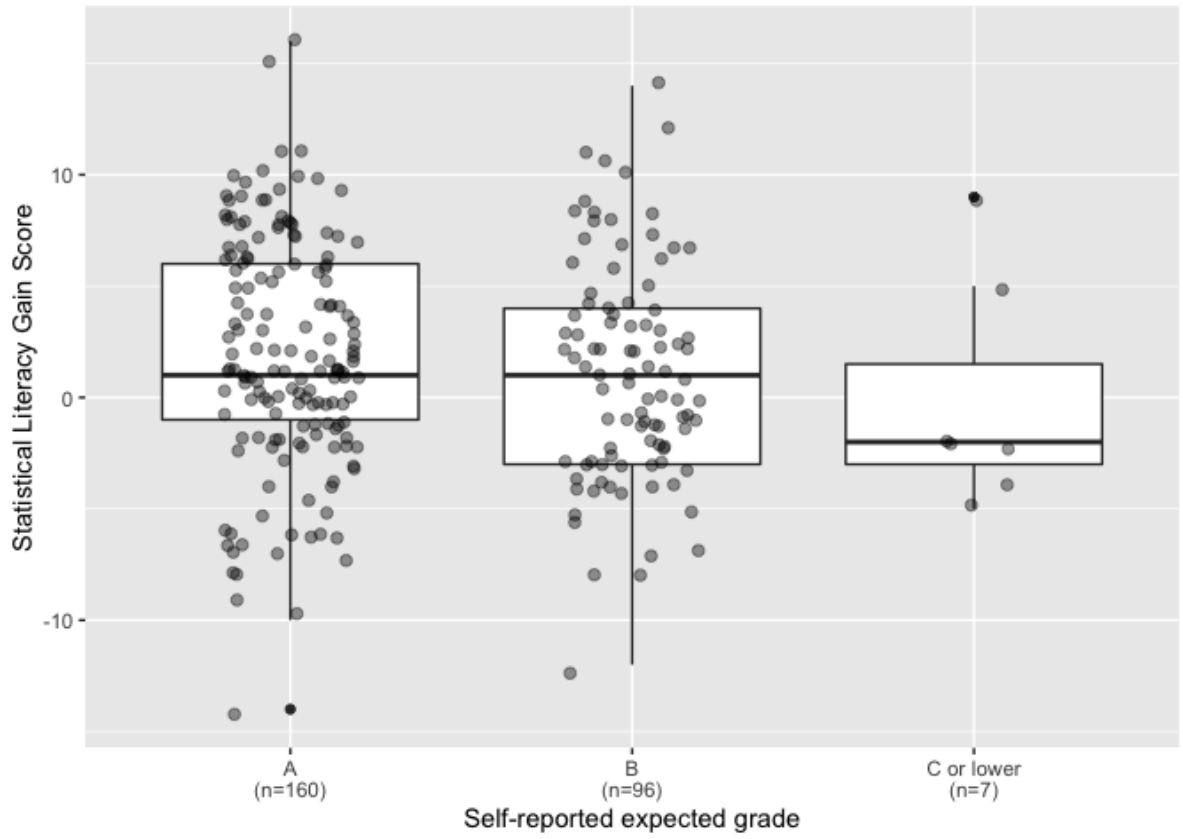
**Figure C.4.** Boxplot of gain score by class standing



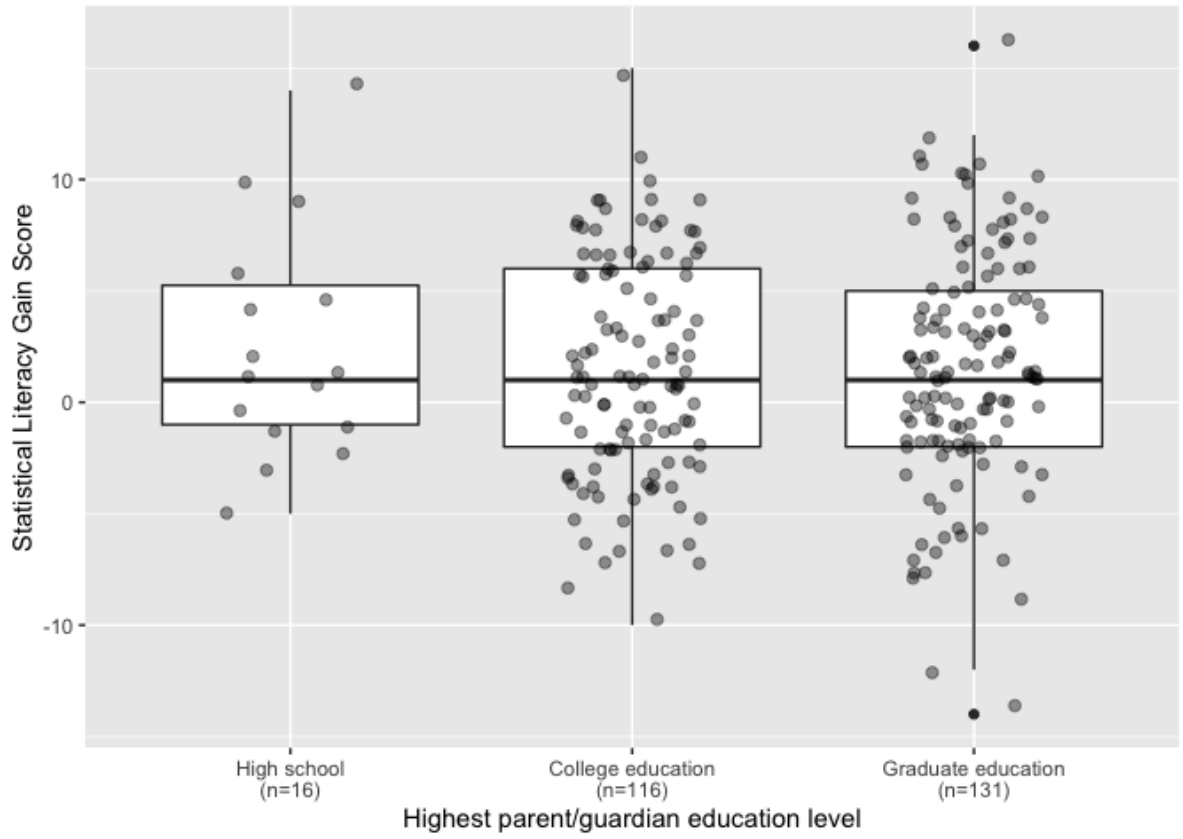
**Figure C.5.** Boxplot of gain score by prior statistics training



**Figure C.6.** Boxplot of gain score by whether a respondent is an international student

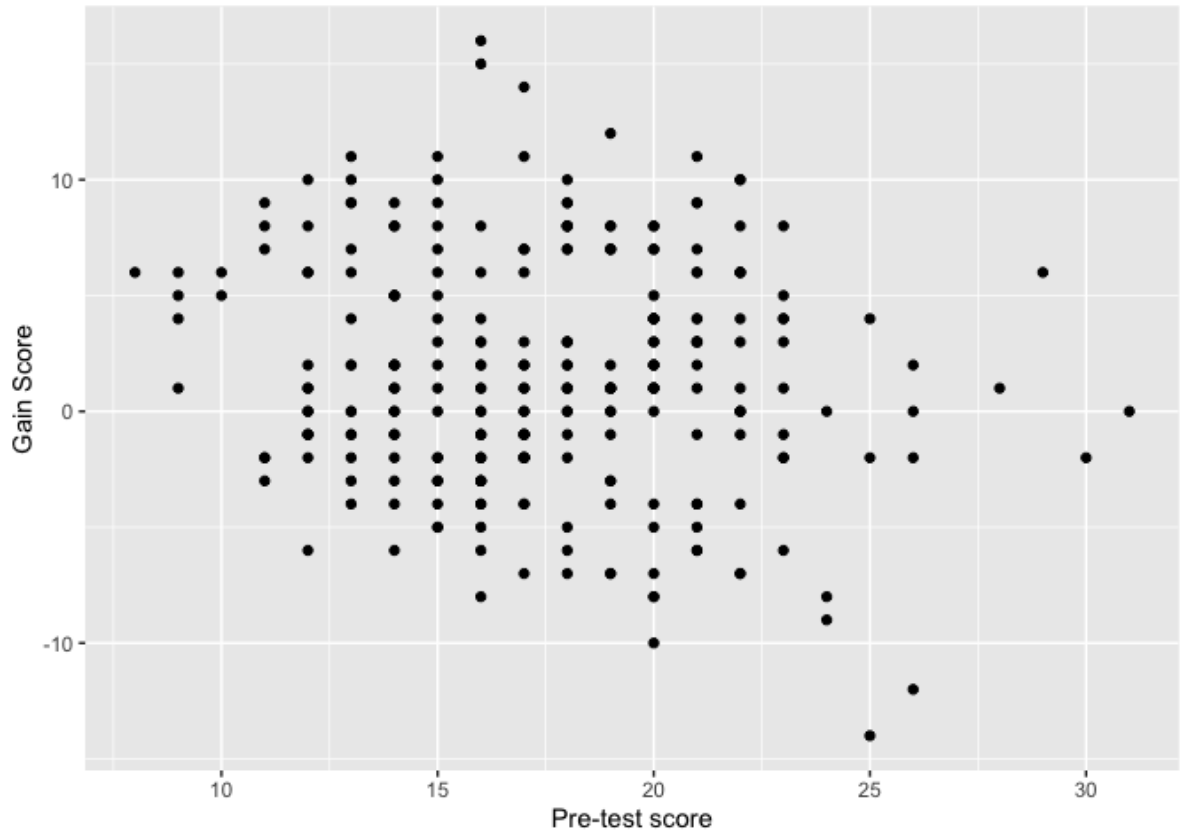


**Figure C.7.** Boxplot of gain score by self-reported expected course grade

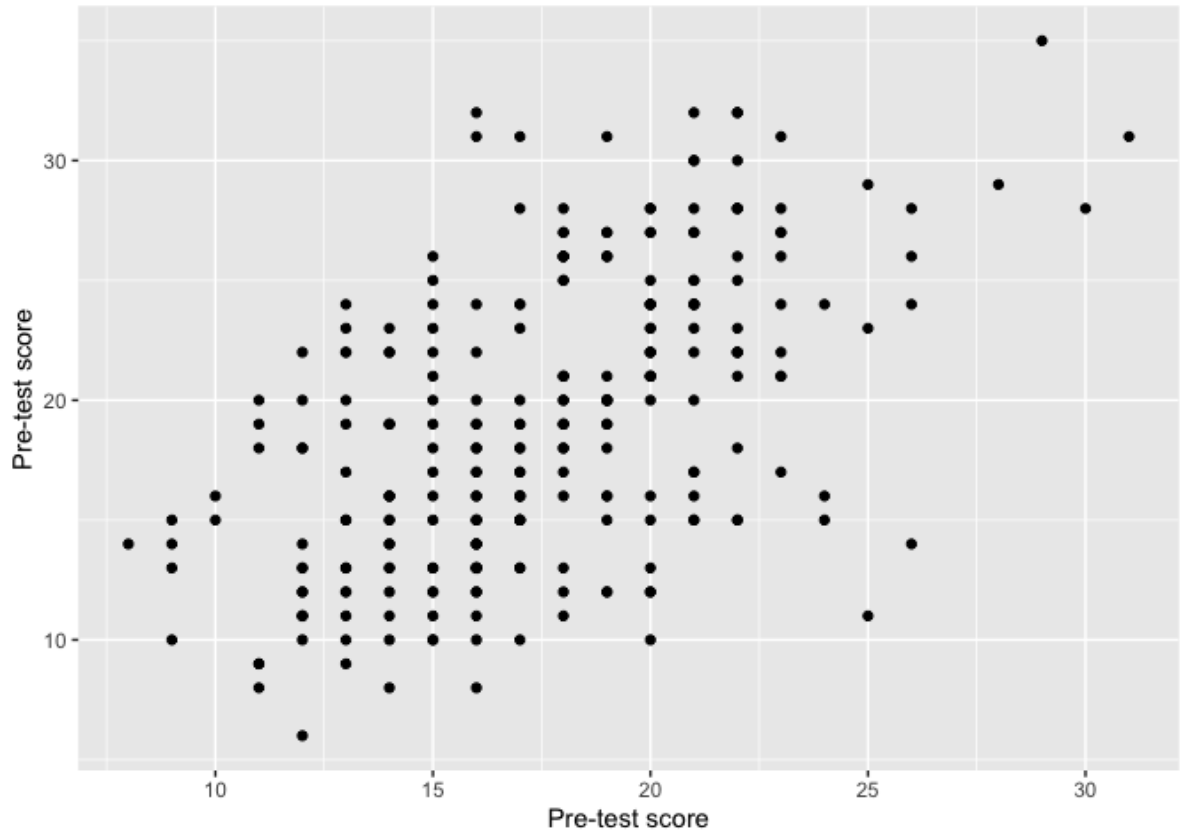


**Figure C.8.** Boxplot of gain score by highest education level of a parent/guardian





**Figure C.9.** Scatterplot of gain score and pre-test score



**Figure C.10.** Scatterplot of post-test and pre-test scores

# Appendix D | Q-matrix for Statistical Literacy

Full descriptions of the skills listed in the Q-matrix are as follows:

- *CommunicateInterpret*: Communicate/interpret statistical results.
- *Descriptive*: Answer a statistical question based on descriptive statistics.
- *Inferential*: Answer a statistical question based on inferential statistics.
- *Visualizations*: Answer a statistical question based on visualizations.
- *Univariate*: Answer a question based on univariate statistics/information.
- *Bivariate*: Answer a question based on bivariate statistics/information.
- *StudyDesign*: Understand study design in order to answer a statistical question.
- *ContextCOVID*: Be familiar with the context - COVID-19 - an item pertains to.

	Communicate	Interpret	Descriptive	Inferential	Visualizations	Univariate	Bivariate	StudyDesign	ContextCOVID
Item 1	1	1	1	0	0	1	0	0	1
Item 2	1	1	1	0	0	1	0	1	1
Item 3	0	1	1	0	1	1	0	1	1
Item 4	0	0	0	0	0	0	1	1	1
Item 5	0	0	0	0	0	1	0	0	1
Item 6	0	0	0	0	0	0	1	0	1
Item 7	1	1	1	0	0	1	0	0	1
Item 8	1	1	1	0	0	1	0	1	1
Item 9	1	1	1	0	1	1	0	0	1
Item 10	1	1	1	0	1	1	0	0	1
Item 11	1	1	1	1	0	0	1	0	1
Item 12	1	1	1	0	0	1	0	0	1
Item 13	1	1	1	0	0	1	0	0	0
Item 14	0	1	1	0	1	1	0	1	1
Item 15	1	1	1	0	0	1	0	0	1
Item 16	1	1	1	0	0	1	0	0	0
Item 17	1	1	1	0	1	1	0	1	0
Item 18	1	1	1	1	1	1	0	1	1
Item 19	1	1	0	1	1	1	0	1	1
Item 20	1	1	0	1	0	1	0	0	1
Item 21	1	1	0	1	0	1	0	0	1
Item 22	0	1	0	1	1	1	0	0	1
Item 23	1	1	0	1	1	0	1	0	0

	Communicate	Interpret	Descriptive	Inferential	Visualizations	Univariate	Bivariate	StudyDesign	ContextCOVID
Item 24	1	0	1	1	1	0	1	1	0
Item 25	1	0	1	1	1	0	1	0	1
Item 26	0	0	1	1	1	0	1	0	1
Item 27	1	0	1	1	0	1	0	0	1
Item 28	1	1	1	1	0	0	1	0	1
Item 29	1	0	1	1	0	0	1	0	1
Item 30	1	0	1	1	0	0	1	0	1
Item 31	0	0	1	1	0	0	0	0	0
Item 32	1	0	1	1	0	0	1	0	0
Item 33	1	0	1	1	0	1	0	1	1
Item 34	0	0	0	0	0	0	1	1	1
Item 35	1	0	0	0	0	0	0	1	1
Item 36	0	1	1	0	1	0	1	0	1
Item 37	1	0	1	1	1	0	1	0	1

Table D.1: Q-matrix for Statistical Literacy using MB-LIS/BLIS

# Bibliography

- [1] BEN-ZVI, D., K. MAKAR, and J. GARFIELD (eds.) (2018) *International Handbook of Research in Statistics Education*, Springer.  
URL <http://link.springer.com/10.1007/978-3-319-66195-7>
- [2] BEN-ZVI, D. and J. GARFIELD (eds.) (2004) *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*, Kluwer Academic Publishers.  
URL <http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf>
- [3] BEN-ZVI, D. and J. GARFIELD (2008) “Introducing the Emerging Discipline of Statistics Education,” *School Science and Mathematics*, **108**(8), pp. 355–361.
- [4] ENGEL, J. (2017) “Statistical literacy for active citizenship: A call for data science education,” *Statistics Education Research Journal*, **16**(1), pp. 44–49.
- [5] GAL, I. (2002) “Adults’ statistical literacy: Meanings, components, responsibilities,” *International Statistical Review*, **70**(1), pp. 1–25.
- [6] GARFIELD, J., R. DEL MAS, and A. ZIEFFLER (2010) “Assessing important learning outcomes in introductory tertiary statistics courses,” *Assessment methods in statistical education: An international perspective*, pp. 75–86.
- [7] GOULD, R. (2017) “Data literacy is statistical literacy,” *Statistics Education Research Journal*, **16**(1), pp. 22–25.
- [8] RUMSEY, D. J. (2002) “Statistical literacy as a goal for introductory statistics courses,” *Journal of Statistics Education*, **10**(3).
- [9] SCHIELD, M. (1999) “Statistical literacy: Thinking critically about statistics,” *Of Significance*, **1**(1), pp. 0–7.  
URL <http://web.augsburg.edu/~mschield/MiloPapers/984StatisticalLiteracy6.pdf>
- [10] UTTS, J. (2021) “Enhancing Data Science Ethics Through Statistical Education and Practice,” *International Statistical Review*, **89**(1), pp. 1–17.
- [11] WALLMAN, K. K. (1993) “Enhancing Statistical Literacy: Enriching Our Society,” *Journal of the American Statistical Association*, **88**(421), pp. 1–8.

- [12] WATSON, J. M. (1998) “The role of statistical literacy in decisions about risk: Where to start,” *For the Learning of Mathematics*, **18**(3), pp. 25–27.
- [13] WATSON, J. M. and R. CALLINGHAM (2003) “Statistical literacy: A complex hierarchical construct,” *Statistics Education Research Journal*, **2**(2), pp. 3–46.  
URL [http://www.stat.auckland.ac.nz/~7B{~}{%}7Diase/serj/SERJ2\(2\){\\_}Watson{\\_%}Callingham.pdf](http://www.stat.auckland.ac.nz/~7B{~}{%}7Diase/serj/SERJ2(2){_}Watson{_%}Callingham.pdf)
- [14] WEILAND, T. (2017) “Problematizing statistical literacy: An intersection of critical and statistical literacies,” *Educational Studies in Mathematics*, **96**(1), pp. 33–47.
- [15] WILD, C. J. (2017) “Statistical literacy as the earth moves,” *Statistics Education Research Journal*, **16**(1), pp. 31–37.
- [16] WILD, C. J. and M. PFANNKUCH (1999) “Statistical Thinking in Empirical Enquiry,” *International Statistical Review*, **67**(3), pp. 223–265.
- [17] ZIEGLER, L. (2014) *Reconceptualizing Statistics Literacy: Developing an Assessment for the Modern Introductory Statistics Course*, Ph.D. thesis, University of Minnesota, <http://hdl.handle.net/11299/165153>.  
URL <http://hdl.handle.net/11299/165153>
- [18] GAISE COLLEGE REPORT ASA REVISION COMMITTEE (2016) *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016, Tech. rep.*, American Statistical Association.  
URL <http://www.amstat.org/education/gaise>.
- [19] PEARL, D. K., J. B. GARFIELD, R. C. DELMAS, R. E. GROTH, J. J. KAPLAN, H. MCGOWAN, and H. S. LEE (2012) *Connecting research to practice in a culture of assessment for introductory college-level statistics, Tech. rep.*, American Statistical Association.
- [20] COMMITTEE, G. P.-K. R. (2020) *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II) - A Framework for Statistics and Data Science Education, Tech. rep.*, National Council of Teachers of Mathematics, [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3).
- [21] ASA, “Education, American Statistical Association,” .  
URL <https://www.amstat.org/ASA/Education/home.aspx>
- [22] PARIS21, “PARIS21 Partnership,” .  
URL <https://paris21.org>
- [23] SHARMA, S. (2017) “Definitions and models of statistical literacy: a literature review,” *Open Review of Educational Research*, **4**(1), pp. 118–133.

- [24] WILKS, S. S. (1951) “Presidential Address,” *Journal of the American Statistical Association*, **46**(253), pp. 1–18.  
URL <https://www.causeweb.org/cause/resources/library/r1266/>
- [25] WATSON, J. and R. CALLINGHAM (2020) “COVID-19 and the need for statistical literacy,” *Australian Mathematics Education Journal (AMEJ)*, **2**(2).
- [26] RATNAWATI, O. A., T. Y. E. SISWONO, and P. WIJAYANTI (2020) “Statistical Literacy Comprehension of Students in the Context of COVID-19 with Collaborative Problem Solving (CPS),” *Math Didactic: Jurnal Pendidikan Matematika*, **6**(3), pp. 264–276.
- [27] BARBIERI, G. A. and P. GIACCHÉ (2006) “The Worth of Data: The Tale of an Experience for Promoting and Improving Statistical Literacy,” *Icots 7 2006*, pp. 1–6.  
URL <http://iase-web.org/documents/papers/icots7/1A1{ }BARB.pdf>
- [28] CARMICHAEL, C. S. (2010) *The development of middle school children’s interest in statistical literacy*, Ph.D. thesis, University of Tasmania.
- [29] FERLIGOJ, A. (2015) “How to improve statistical literacy?” *Metodoloski Zvezki*, **12**(1), pp. 1–10.
- [30] SCHIELD, M. (2004) “Statistical literacy curriculum design,” *Curricular Development in Statistics Education*, pp. 54–74.  
URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.144.8102{&}rep=rep1{&}type=pdf>
- [31] SUHERMI and D. B. WIDJAJANTI (2020) “What are the roles of technology in improving student statistical literacy?” *Journal of Physics: Conference Series*, **1581**.
- [32] WATSON, J. M. (2011) “Foundations for Improving Statistical Literacy,” *Statistical Journal of the IAOS*, **27**, pp. 197–204.  
URL [10.3233/SJI-2011-0728](https://doi.org/10.3233/SJI-2011-0728)
- [33] SCHIELD, M. (2017) “GAISE 2016 promotes statistical literacy,” *Statistics Education Research Journal*, **16**(1), pp. 50–54.
- [34] NEUMANN, D. L., M. HOOD, and M. M. NEUMANN (2013) “Using real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course,” *Statistics Education Research Journal*, **12**(2), pp. 59–70.
- [35] RAO, C. R. (1975) “Teaching of statistics at the secondary level An interdisciplinary approach,” *International Journal of Mathematical Education in Science and Technology*, **6**(2), pp. 151–162.



- [36] ZIEFFLER, A., J. GARFIELD, S. ALT, D. DUPUIS, K. HOLLEQUE, and B. CHANG (2008) “What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature,” *Journal of Statistics Education*, **16**(2).  
URL <https://doi.org/10.1080/10691898.2008.11889566>
- [37] BURRILL, G. and M. CAMDEN (2004) *Curricular Development in Statistics Education, Tech. rep.*, International Association for Statistical Education (IASE), Lund, Sweden.
- [38] HALL, M. R. and G. H. ROWELL (2008) “Introductory Statistics Education and the National Science Foundation,” *Journal of Statistics Education*, **16**(2).
- [39] (ASA), A. S. A. (2005) *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2005, Tech. rep.*  
URL [https://www.amstat.org/docs/default-source/amstat-documents/2005gaisecollege\\_{\\_}full.pdf](https://www.amstat.org/docs/default-source/amstat-documents/2005gaisecollege_{_}full.pdf)
- [40] ZIEFFLER, A., J. GARFIELD, R. C. DELMAS, L. LE, R. ISAAK, A. BJORNS-DOTTIR, and J. PARK (2011) “Publishing in SERJ: An analysis of papers from 2002-2009,” *Statistics Education Research Journal*, **10**(2), pp. 5–26.
- [41] WATSON, J. (2016) “Whither Statistics Education Research?” in *Mathematics Education Research Group of Australasia*, Adelaide, pp. 33–58.  
URL <https://eric.ed.gov/?q=statistics+AND+education+AND+research{%&}id=ED572360>
- [42] TISHKOVSKAYA, S. and G. A. LANCASTER (2012) “Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform,” *Journal of Statistics Education*, **20**(2).
- [43] SCHWAB-MCCOY, A. (2019) “The State of Statistics Education Research in Client Disciplines: Themes and Trends Across the University,” *Journal of Statistics Education*, **27**(3), pp. 253–264.  
URL <https://doi.org/10.1080/10691898.2019.1687369>
- [44] DATA DESCRIPTION, “Data and Story Library (DASL),” .  
URL <https://dasl.datadescription.com>
- [45] GAL, I. (2019) “Understanding statistical literacy: About knowledge of contexts and models,” *Actas del tercer Congreso Internacional Virtual de Educación Estadística*.  
URL <http://digibug.ugr.es/bitstream/handle/10481/55029/gal.pdf?sequence=1{%&}isAllowed=y>
- [46] GARFIELD, J., R. DELMAS, and A. ZIEFFLER (2012) “Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course,” *ZDM - International Journal on Mathematics Education*, **44**(7), pp. 883–898.

- [47] LEE, H. and D. TRAN (2015) “Statistical habits of mind,” *Teaching statistics through data investigations MOOC-Ed*, pp. 1–4.
- [48] KOKLU, O. (2017) *Undergraduate Students’ Informal Notions of Variability*, Ph.D. thesis, University of Georgia.
- [49] PFANNKUCH, M. (2011) “The role of context in developing informal statistical inferential reasoning: A classroom study,” *Mathematical Thinking and Learning*, **13**(1-2), pp. 27–46.
- [50] BROWN, K. M. (2019) “More Questions and Fewer Contexts: Designing Exercises for Statistics Courses,” *Journal of Statistics Education*, **27**(3), pp. 216–224.  
URL <https://doi.org/10.1080/10691898.2019.1669508>
- [51] WATSON, J. (2015) “Statistical literacy in action,” *Australian Primary Mathematics Classroom*, **20**(4), pp. 26–30.  
URL <http://libezproxy.open.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=111671346&site=ehost-live&scope=site>
- [52] YOLCU, A. (2012) *An Investigation of Eighth Grade Students’ Statistical Literacy, Attitudes Towards Statistics and Their Relationship*, Ph.D. thesis, Middle East Technical University.
- [53] ——— (2014) “Middle school students’ statistical literacy: Role of grade level and gender,” *Statistics Education Research Journal*, **13**(2), pp. 118–131.
- [54] WATSON, J. M. and B. A. KELLY (1993) “The Vocabulary of Statistical Literacy,” *Educational Research, Risks, and Dilemmas: Proceedings of the Joint Conferences of the New Zealand Association for Research in Education and the Australian Association for Research in Education*.
- [55] NORTH, D., I. GAL, and T. ZEWOTIR (2014) “Building capacity for developing statistical literacy in a developing country: Lessons learned from an intervention,” *Statistics Education Research Journal*, **13**(2), pp. 15–27.
- [56] CONTI, K. C. and D. L. DE CARVALHO (2014) “Statistical literacy: Developing a youth and adult education statistical project,” *Statistics Education Research Journal*, **13**(2), pp. 164–176.
- [57] GRANT, R. (2017) “Statistical literacy in the data science workplace,” *Statistics Education Research Journal*, **16**(1), pp. 17–21.
- [58] SUTHERLAND, S. and J. RIDGWAY (2017) “Interactive visualisations and statistical literacy,” *Statistics Education Research Journal*, **16**(1), pp. 26–30.

- [59] TUNSTALL, S. L. (2018) “Investigating College Students’ Reasoning With Messages of Risk and Causation,” *Journal of Statistics Education*, **26**(2), pp. 76–86.  
URL <https://doi.org/10.1080/10691898.2018.1456989>
- [60] BUDGETT, S. and M. PFANKUCH (2007) “Assessing Students’ Statistical Literacy,” in *IASE/ISI Satellite*, pp. 1–7.
- [61] KAPLAN, J. J. and J. THORPE (2010) “Post secondary and adult statistical literacy: Assessing beyond the classroom,” *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8)*, **8**.
- [62] SABBAG, A., J. GARFIELD, and A. ZIEFFLER (2018) “Assessing statistical literacy and statistical reasoning: The REALI instrument,” *Statistics Education Research Journal*, **17**(2), pp. 141–160.
- [63] SANCHEZ, J. (2007) “Building Statistical Literacy Assessment Tools With the IASE/ISLP,” *IASE/ISI Satellite*.  
URL <https://iase-web.org/documents/papers/sat2007/Sanchez.pdf>
- [64] ZIEGLER, L. and J. GARFIELD (2018) “Developing a statistical literacy assessment for the modern introductory statistics course,” *Statistics Education Research Journal*, **17**(2), pp. 161–178.
- [65] COUNCIL, N. R. (2000) *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*, The National Academies Press, Washington, DC.  
URL <https://doi.org/10.17226/9853>
- [66] GAL, I. (2003) “Teaching for statistical literacy and services of statistics agencies,” *American Statistician*, **57**(2), pp. 80–84.
- [67] LOVETT, M. C. and J. R. GREENHOUSE (2000) “Applying Cognitive Theory to Statistics Instruction,” *American Statistician*, **54**(3), pp. 196–206.
- [68] BECKMAN, M. D. (2015) *Assessment of cognitive transfer outcomes for students of introductory statistics*, Ph.D. thesis, University of Minnesota, <http://hdl.handle.net/11299/175709>.  
URL <https://conservancy.umn.edu/handle/11299/175709>
- [69] PAAS, F. G. W. C. (1992) “Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach,” *Journal of Educational Psychology*, **84**(4), pp. 429–434.
- [70] PERKINS, D. N. and G. SALOMON (1992) “Transfer of learning,” in *International Encyclopedia of Education*, Pergamon Press, Oxford, England.

- [71] BONOTTO, C. (2002) "Suspension of sense-making in mathematical word problem solving: A possible remedy," in *Proceedings of the 2nd International Conference on the Teaching of Mathematics*.
- [72] GREER, B., L. VERSCHAFFEL, and S. MUKHOPADHYAY (2007) "Modelling for Life: Mathematics and Children's Experience," in *Modeling and Applications in Mathematics Education* (W. Blum, P. L. Galbraith, H.-W. Henn, and M. Niss, eds.), 14 ed., chap. 2.6, Springer, pp. 89–98.
- [73] GICK, M. L. and K. J. HOLYOAK (1980) "Analogical problem solving," *Cognitive Psychology*, **12**(3), pp. 306–355.
- [74] MILLAR, R. and S. MANOHARAN (2021) "Repeat individualized assessment using isomorphic questions: a novel approach to increase peer discussion and learning," *International Journal of Educational Technology in Higher Education*, **18**(1). URL <https://doi.org/10.1186/s41239-021-00257-y>
- [75] WILLIAMSON, D. M., M. S. JOHNSON, S. SINHARAY, and I. I. BEJAR (2002) *Hierarchical IRT Examination of Isomorphic Equivalence of Complex Structured Response Tasks, Tech. rep.*, Educational Testing Service, Princeton, NJ.
- [76] LEHRER, R. and L. SCHAUBLE (2007) "A Developmental Approach for Supporting the Epistemology of Modeling," in *Modeling and Applications in Mathematics Education* (W. Blum, P. L. Galbraith, H.-W. Henn, and M. Niss, eds.), 14 ed., chap. 3.1.4, Springer, pp. 153–160.
- [77] FAY, D. M., R. LEVY, and V. MEHTA (2018) "Investigating Psychometric Isomorphism for Traditional and Performance-Based Assessment," *Journal of Educational Measurement*, **55**(1), pp. 52–77.
- [78] BARNIOL, P. and G. ZAVALA (2014) "Force, velocity, and work: The effects of different contexts on students' understanding of vector concepts using isomorphic problems," *Physical Review Special Topics - Physics Education Research*, **10**(2), pp. 1–15.
- [79] KUSAIRI, S., A. HIDAYAT, and N. HIDAYAT (2017) "Web-based Diagnostic Test: Introducing Isomorphic Items to Assess Students' Misconceptions and Error Patterns," *Chemistry: Bulgarian Journal of Science Education*, **26**(4).
- [80] KUSAIRI, S., D. A. PUSPITA, A. SURYADI, and H. SUWONO (2020) "Physics Formative Feedback Game: Utilization of Isomorphic Multiple-choice Items to Help Students Learn Kinematics," *TEM Journal*, **9**(4), pp. 1625–1632.
- [81] LIN, S.-Y. Y. and C. SINGH (2011) "Using isomorphic problems to learn introductory physics," *Physical Review Special Topics - Physics Education Research*, **7**(2), p. 20104.

- [82] LUGER, G. F. and M. A. BAUER (1978) “Transfer effects in isomorphic problem situations,” *Acta Psychologica*, **42**(2), pp. 121–131.
- [83] SUGANDA, T., S. KUSAIRI, N. AZIZAH, and P. PARNO (2020) “The Correlation of Isomorphic, Open-Ended, and Conventional Score on the Ability to Solve Kinematics Graph Questions,” *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, **6**(2), pp. 173–180.
- [84] PARKER, M. C., M. GUZDIAL, and S. ENGLEMAN (2016) “Replication, validation, and use of a language independent CS1 knowledge assessment,” *ICER 2016 - Proceedings of the 2016 ACM Conference on International Computing Education Research*, pp. 93–101.
- [85] BASSOK, M. and K. J. HOLYOAK (1989) “Interdomain Transfer Between Isomorphic Topics in Algebra and Physics,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**(1), pp. 153–166.
- [86] COOK, D. A. and R. HATALA (2016) “Validation of educational assessments: a primer for simulation and beyond,” *Advances in Simulation*, **1**(31), pp. 1–12.  
URL <http://dx.doi.org/10.1186/s41077-016-0033-y>
- [87] COBB, G. W. (2007) “The Introductory Statistics Course: A Ptolemaic Curriculum,” *Technology Innovations in Statistics Education*, **1**(1), pp. 1–16.  
URL <http://www.escholarship.org/uc/item/6hb3k0nz%5Cnpapers3://publication/uuid/426B647B-BAD9-4278-8F02-990BD513D5FA>
- [88] MOSTELLER, F. and R. F. BORUCH (eds.) (2002) *Evidence matters: Randomized trials in education research*, Brookings Institution Press.
- [89] MOTZ, B. A., P. F. CARVALHO, J. R. DE LEEUW, and R. L. GOLDSTONE (2018) “Embedding Experiments: Staking Causal Inference in Authentic Educational Contexts,” *Journal of Learning Analytics*, **5**(2), pp. 47–59.
- [90] RAUDENBUSH, S. W. and D. SCHWARTZ (2020) “Randomized experiments in education, with implications for multilevel causal inference,” *Annual Review of Statistics and Its Application*, **7**, pp. 177–208.
- [91] COOK, T. D. (2002) “Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them,” *Educational Evaluation and Policy Analysis*, **24**(3), pp. 175–199.
- [92] CORDERO, J. M., V. CRISTÓBAL, and D. SANTÍN (2018) “Causal Inference on Education Policies: a Survey of Empirical Studies Using Pisa, Timss and Pirls,” *Journal of Economic Surveys*, **32**(3), pp. 878–915.

- [93] WEISSMAN, M. B. (2021) “Policy recommendations from causal inference in physics education research,” *Physical Review Physics Education Research*, **17**(2), p. 20118.  
URL <https://doi.org/10.1103/PhysRevPhysEducRes.17.020118>
- [94] IMBENS, G. W. (2011) “Experimental Design for Unit and Cluster Randomized Trials,” *3Ie Working Paper*, (June).
- [95] MAURER, K. and D. LOCK (2016) “Comparison of Learning Outcomes for Simulation-based and Traditional Inference Curricula in a Designed Educational Experiment,” *Technology Innovations in Statistics Education*, **9**(1).  
URL <https://escholarship.org/uc/item/0wm523b0>
- [96] LESSER, L. M., D. K. PEARL, and J. J. WEBER (2016) “Assessing fun items’ effectiveness in increasing learning of college introductory statistics students: Results of a randomized experiment,” *Journal of Statistics Education*, **24**(2), pp. 54–62.
- [97] DIMELLA, T. (2019) *Teaching Statistics With a Critical Pedagogy*, Ph.D. thesis, Appalachian State University.
- [98] WANG, S. L., A. Y. ZHANG, S. MESSER, A. WIESNER, and D. K. PEARL (2021) “Student-Developed Shiny Applications for Teaching Statistics,” *Journal of Statistics and Data Science Education*, **29**(3), pp. 218–227.  
URL <https://doi.org/10.1080/26939169.2021.1995545>
- [99] ARTIST (2006), “ARTIST topic scales,” .  
URL [apps3.cehd.umn.edu/artist/tests/index.html](http://apps3.cehd.umn.edu/artist/tests/index.html)
- [100] SCHAU, C., J. STEVENS, T. L. DAUPHINEE, and A. D. VECCHIO (1995) “The Development and Validation of the Survey of Attitudes Toward Statistics,” *Educational and Psychological Measurement*, **55**(5).
- [101] EARP, M. S. (2007) *Development and validation of the Statistics Anxiety Measure*, phdthesis, University of Denver.  
URL <http://login.ezproxy.lib.umn.edu/login?url=http://search.proquest.com/docview/304862597?accountid=14586%}5Cnhttp://primo.lib.umn.edu/openurl/TWINCITIES/TWINCITIES{ }SP?url{ }ver=Z39.88-2004{&}rft{ }val{ }fmt=info:ofi/fmt:kev:mtx:dissertation{&}genre=dissertations+{&}+th>
- [102] DELMAS, R., J. GARFIELD, A. OOMS, and B. CHANCE (2007) “Assessing Students’ Conceptual Understanding After a First Course in Statistics,” *Statistics Education Research Journal*, **6**(2), pp. 28–58.
- [103] TAYLOR, P. C., B. J. FRASER, and D. L. FISHER (1997) “Monitoring constructivist classroom learning environments,” *International Journal of Educational Research*, **27**(4), pp. 293–302.

- [104] STEFFE, L. P. and P. W. THOMPSON (2000) “Teaching experiment methodology: Underlying principles and essential elements,” in *Research design in mathematics and science education* (R. Lesh and A. E. Kelly, eds.), Erlbaum, Hillsdale, NJ, pp. 267–307.
- [105] SALDANHA, L. A. and P. W. THOMPSON (2007) “Exploring Connections between Sampling Distributions and Statistical Inference: an Analysis of Students’ Engagement and Thinking in the Context of Instruction Involving Repeated Sampling,” *International Electronic Journal of Mathematics Education*, **2**(3), pp. 270–297.
- [106] KAPLAN, J. J., N. T. ROGNESS, and D. G. FISHER (2014) “Exploiting lexical ambiguity to help students understand the meaning of Random,” *Statistics Education Research Journal*, **13**(1), pp. 9–24.
- [107] STOKER, G. and P. JOHN (2009) “Design experiments: Engaging policy makers in the search for evidence about what works,” *Political Studies*, **57**(2), pp. 356–373.
- [108] MVUDUDU, N. (2003) “A cross-cultural study of the connection between students’ attitudes toward statistics and the use of constructivist strategies in the course,” *Journal of Statistics Education*, **11**(3).
- [109] XU, C., M. PETERS, and S. BROWN (2020) “Instructor and Instructional Effects on Students’ Statistics Attitudes,” *Statistics Education Research Journal*, **19**(2), pp. 7–26.
- [110] HONG, G. and S. W. RAUDENBUSH (2008) “Causal inference for time-varying instructional treatments,” *Journal of Educational and Behavioral Statistics*, **33**(3), pp. 333–362.
- [111] KAPLAN, D. (2016) “Causal inference with large-scale assessments in education from a Bayesian perspective: a review and synthesis,” *Large Scale Assessment Education*.  
URL doi:10.1186/s40536-016-0022-6
- [112] YAO, L., Z. CHU, S. LI, Y. LI, J. GAO, and A. ZHANG (2021) “A Survey on Causal Inference,” *ACM Transactions on Knowledge Discovery from Data*, **15**(5), 2002.02770.
- [113] EIDE, E. R. and M. H. SHOWALTER (2012) “Methods matter: Improving causal inference in educational and social science research: A review article,” *Economics of Education Review*, **31**, pp. 744–748.  
URL <http://dx.doi.org/10.1016/j.econedurev.2012.05.010>
- [114] RUBIN, D. B. (1974) “Estimating causal effects of treatment in randomized and nonrandomized studies,” *Journal of Educational Psychology*, **66**(5), pp. 688–701.  
URL <http://www.fsb.muohio.edu/lij14/420{ }paper{ }Rubin74.pdf>

- [115] ——— (2005) “Causal inference using potential outcomes: Design, modeling, decisions,” *Journal of the American Statistical Association*, **100**(469), pp. 322–331.
- [116] ——— (1980) “Randomization analysis of experimental data: The fisher randomization test,” *Journal of the American Statistical Association*, **75**(371), pp. 575–582.
- [117] DE AYALA, R. J. (2013) *The Theory and Practice of Item Response Theory*, Guilford Publications.
- [118] CHALMERS, R. P. (2012) “Mirt: A multidimensional item response theory package for the R environment,” *Journal of Statistical Software*, **48**(6).
- [119] TATSUOKA, K. K. (1983) “Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory,” *Journal of Educational Measurement*, **20**(4), pp. 345–354.
- [120] DE LA TORRE, J. and C. Y. CHIU (2016) “A General Method of Empirical Q-matrix Validation,” *Psychometrika*, **81**(2), pp. 253–273.
- [121] BIRENBAUM, M. and K. K. TATSUOKA (1993) “Applying an IRT-Based Cognitive Diagnostic Model to Diagnose Students’ Knowledge States in Multiplication and Division with Exponents,” *Applied Measurement in Education*, **6**(4), pp. 255–268.
- [122] DE LA TORRE, J. (2011) “The Generalized DINA Model Framework,” *Psychometrika*, **76**(3), pp. 510–510.  
URL <http://link.springer.com/10.1007/s11336-011-9214-8>
- [123] HAERTEL, E. H. (1989) “Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items,” *Journal of Educational Measurement*, **26**(4), pp. 301–321.
- [124] VON DAVIER, M. and K. YAMAMOTO (2004) “A Class of Models for Cognitive Diagnosis,” in *Fourth Spearman Conference*, Philadelphia, PA.
- [125] HENSON, R. A., J. L. TEMPLIN, and J. T. WILLSE (2009) “Defining a family of cognitive diagnosis models using log-linear models with latent variables,” *Psychometrika*, **74**(2), pp. 191–210.
- [126] VON DAVIER, M. (2014) “The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM),” *ETS Research Report Series*, **2014**(2), pp. 1–13.
- [127] MA, W. and J. DE LA TORRE (2020) “GDINA : An R Package for Cognitive Diagnosis Modeling,” *Journal of Statistical Software*, **93**(14), pp. 1–26.  
URL <http://www.jstatsoft.org/v93/i14/>



- [128] DE LA TORRE, J. (2009) “DINA model and parameter estimation: A didactic,” *Journal of Educational and Behavioral Statistics*, **34**(1), pp. 115–130.
- [129] TEMPLIN, J. L. and R. A. HENSON (2006) “Measurement of psychological disorders using cognitive diagnosis models,” *Psychological Methods*, **11**(3), pp. 287–305.
- [130] JUNKER, B. W. and K. SIJTSMA (2001) “Cognitive Assessment Models With Few Assumptions, and Connections With Nonparametric Item Response Theory,” *Applied Psychological Measurement*, **25**(3), pp. 258–272.
- [131] MARIS, E. (1999) “Estimating Multiple Classification Latent Class Models,” *Psychometrika*, **64**(2), pp. 187–212.
- [132] HARTZ, S. M. (2002) “A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality,” *APA PsycInfo*, pp. 1–168.
- [133] DEMPSTER, A. P., N. M. LAIRD, and D. B. RUBIN (1977) “Maximum Likelihood from Incomplete Data Via the EM Algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), pp. 1–22.
- [134] YI, Y. S. (2017) “Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models,” *Language Testing*, **34**(3), pp. 337–355.
- [135] ——— (2017) “In Search of Optimal Cognitive Diagnostic Model(s) for ESL Grammar Test Data,” *Applied Measurement in Education*, **30**(2), pp. 82–101.
- [136] YAMAGUCHI, K. and K. OKADA (2018) “Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment,” *PLoS ONE*, **13**(2).
- [137] HU, J., M. D. MILLER, A. C. HUGGINS-MANLEY, and Y. H. CHEN (2016) “Evaluation of Model Fit in Cognitive Diagnosis Models,” *International Journal of Testing*, **16**(2), pp. 119–141.
- [138] SORREL, M. A., F. J. ABAD, J. OLEA, J. DE LA TORRE, and J. R. BARRADA (2017) “Inferential Item-Fit Evaluation in Cognitive Diagnosis Modeling,” *Applied Psychological Measurement*, **41**(8), pp. 614–631.
- [139] SORREL, M. A., J. DE LA TORRE, F. J. ABAD, and J. OLEA (2017) “Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models,” *Methodology*, **13**, pp. 39–47.
- [140] NÁJERA, P., F. J. ABAD, and M. A. SORREL (2021) “Determining the Number of Attributes in Cognitive Diagnosis Modeling,” *Frontiers in Psychology*, **12**(February).

- [141] CULPEPPER, S. A. (2019) “Estimating the Cognitive Diagnosis Q Matrix with Expert Knowledge: Application to the Fraction-Subtraction Dataset,” *Psychometrika*, **84**(2), pp. 333–357.  
URL <https://doi.org/10.1007/s11336-018-9643-8>
- [142] DE LA TORRE, J. (2008) “An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications,” *Journal of Educational Measurement*, **45**(4), pp. 343–362.
- [143] HUANG, H.-Y. and W.-C. WANG (2014) “The Random-Effect DINA Model,” *Journal of Educational Measurement*, **51**(1), pp. 75–97.
- [144] ROBITZSCH, A. and A. C. GEORGE (2014) “Multiple group cognitive diagnosis models, with an emphasis on differential item functioning,” *Psychological Test and Assessment Modeling*, **56**(393-420), pp. 405–432.  
URL <file:///C:/Users/sunpn552/iCloudDrive/IPN/Literature/Robitzsch,George2014-Multiplegroupcognitivediagnosismodels.pdfM4-Citavi>
- [145] PARK, Y. S. and Y. S. LEE (2014) “An extension of the DINA model using covariates: Examining factors affecting response probability and latent classification,” *Applied Psychological Measurement*, **38**(5), pp. 376–390.
- [146] PARK, Y. S., K. XING, and Y. S. LEE (2018) “Explanatory Cognitive Diagnostic Models: Incorporating Latent and Observed Predictors,” *Applied Psychological Measurement*, **42**(5), pp. 376–392.
- [147] DAI, S. and D. S. VALDIVIA (2022) “Dealing with Missing Responses in Cognitive Diagnostic Modeling,” *Psych*, **4**, pp. 318–342.
- [148] KARELITZ, T. M. (2004) *Ordered category attribute coding framework for cognitive assessments*, phdthesis, University of Illinois at Urbana-Champaign.  
URL <https://www.lib.byu.edu/cgi-bin/remoteauth.pl?url=http://search.ebscohost.com/login.aspx?direct=true{%&}db=psych{%&}AN=2005-99009-124{%&}site=ehost-live{%&}scope=site>
- [149] CHEN, J. and J. DE LA TORRE (2013) “A General Cognitive Diagnosis Model for Expert-Defined Polytomous Attributes,” *Applied Psychological Measurement*, **37**(6), pp. 419–437.
- [150] AKAEZE, H. O. (2020) *Incorporating Differential Speed in Cognitive Diagnostic Models with Polytomous Attributes*, phdthesis, Michigan State University.
- [151] HONG, H., C. WANG, Y. S. LIM, and J. DOUGLAS (2015) “Efficient Models for Cognitive Diagnosis With Continuous and Mixed-Type Latent Variables,” *Applied Psychological Measurement*, **39**(1), pp. 31–43.
- [152] STOUT, W. (2007) “Skills Diagnosis Using IRT-Based Continuous Latent Trait Models,” *Journal of Educational Measurement*, **44**(4), pp. 313–324.

- [153] FISCHER, G. H. (1973) “The Linear Logistic Test Model As an Instrument in Educational Research,” *Acta Psychologica*, **37**, pp. 359–374.
- [154] ——— (1983) “Logistic latent trait models with linear constraints,” *Psychometrika*, **48**(1).
- [155] RECKASE, M. D. and R. L. MCKINLEY (1991) “The Discriminating Power of Items That Measure More Than One Dimension,” *Applied Psychological Measurement*, **15**(4), pp. 361–373.
- [156] SYMPSON, J. B. (1978) “A model for testing with multidimensional items,” in *Proceedings of the 1977 computerized adaptive testing conference*.
- [157] EMBRETSON, S. E. (1985) “Multicomponent latent trait models for test design,” in *Test design: Developments in psychology and psychometrics*, Academic Press, pp. 195—218.
- [158] ——— (1997) “Multicomponent response models,” in *Handbook of modern item response theory*, Springer, pp. 305—321.
- [159] WHITELEY, S. E. (1980) “Multicomponent Latent Trait Models for Ability Tests,” *Psychometrika*, **45**(4).
- [160] MINCHEN, N. D., J. DE LA TORRE, and Y. LIU (2017) “A Cognitive Diagnosis Model for Continuous Response,” *Journal of Educational and Behavioral Statistics*, **42**(6), pp. 651–677.
- [161] MINCHEN, N. and J. DE LA TORRE (2018) “A General Cognitive Diagnosis Model for Continuous-Response Data,” *Measurement*, **16**(1), pp. 30–44.
- [162] ZHAN, P., H. JIAO, and D. LIAO (2018) “Cognitive diagnosis modelling incorporating item response times,” *British Journal of Mathematical and Statistical Psychology*, **71**, pp. 262–286.
- [163] FINKELMAN, M. D., J. DE LA TORRE, and J. A. KARP (2020) “Cognitive diagnosis models and automated test assembly: an approach incorporating response times,” *International Journal of Testing*, **20**(4), pp. 299–320.  
URL <https://doi.org/10.1080/15305058.2020.1828427>
- [164] WANG, S. and Y. CHEN (2020) “Using Response Times and Response Accuracy to Measure Fluency Within Cognitive Diagnosis Models,” *Psychometrika*, **85**(3), pp. 600–629.  
URL <https://doi.org/10.1007/s11336-020-09717-2>
- [165] HSU, C.-L., K.-Y. JIN, and M. M. CHIU (2020) “Cognitive Diagnostic Models for Random Guessing Behaviors,” *Frontiers in Psychology*, **11**.

- [166] TATSUOKA, K. K. (1990) “Toward an integration of item-response theory and cognitive error diagnosis,” in *Diagnostic Monitoring of Skill and Knowledge Acquisition* (N. Frederiksen and R. Glaser, eds.), Psychology Press.
- [167] DE LA TORRE, J. and J. A. DOUGLAS (2004) “Higher-order latent trait models for cognitive diagnosis,” *Psychometrika*, **69**(3), pp. 333–353.
- [168] DECARLO, L. T. (2011) “On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix,” *Applied Psychological Measurement*, **35**(1), pp. 8–26.
- [169] DE LA TORRE, J. and Y. S. LEE (2013) “Evaluating the wald test for item-level comparison of saturated and reduced models in cognitive diagnosis,” *Journal of Educational Measurement*, **50**(4), pp. 355–373.
- [170] SESSOMS, J. and R. A. HENSON (2018) “Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary,” *Measurement: Interdisciplinary Research and Perspectives*, **16**(1), pp. 1–17.  
URL <https://doi.org/10.1080/15366367.2018.1435104>
- [171] LEE, Y. S., Y. S. PARK, and D. TAYLAN (2011) “A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007,” *International Journal of Testing*, **11**(2), pp. 144–177.
- [172] GEORGE, A. C. and A. ROBITZSCH (2015) “Cognitive Diagnosis Models in R: A didactic,” *The Quantitative Methods for Psychology*, **11**(3), pp. 189–205.
- [173] WU, X., R. WU, H. H. CHANG, Q. KONG, and Y. ZHANG (2020) “International Comparative Study on PISA Mathematics Achievement Test Based on Cognitive Diagnostic Models,” *Frontiers in Psychology*, **11**(September), pp. 1–13.
- [174] LEE, Y. W. and Y. SAWAKI (2009) “Application of three cognitive diagnosis models to ESL reading and listening assessments,” *Language Assessment Quarterly*, **6**(3), pp. 239–263.
- [175] XU, X. and M. VON DAVIER (2008) *Fitting the Structured General Diagnostic Model To Naep Data, Tech. rep.*, Educational Testing Service.
- [176] LI, H. (2011) “A cognitive diagnostic analysis of the MELAB reading test,” *Spain Fellow Working Papers in Second or Foreign Language Assessment*, **9**(January 2011), pp. 17–46.
- [177] WANG, C. and M. J. GIERL (2011) “Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees’ Cognitive Skills in Critical Reading,” *Journal of Educational Measurement*, **48**(2), pp. 165–187.  
URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2011.00142.x>

- [178] KIM, A. Y. A. (2015) *Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability*, vol. 32.
- [179] LI, H., C. V. HUNTER, and P. W. LEI (2016) “The selection of cognitive diagnostic models for a reading comprehension test,” *Language Testing*, **33**(3), pp. 391–409.
- [180] RAVAND, H. (2016) “Application of a Cognitive Diagnostic Model to a High-Stakes Reading Comprehension Test,” *Journal of Psychoeducational Assessment*, **34**(8), pp. 782–799.
- [181] RAVAND, H. and A. ROBITZSCH (2018) “Cognitive diagnostic model of best choice: a study of reading comprehension,” *Educational Psychology*, **38**(10), pp. 1255–1277.
- [182] HE, L., Z. JIANG, and S. MIN (2021) “Diagnosing writing ability using China’s Standards of English Language Ability: Application of cognitive diagnosis models,” *Assessing Writing*, **50**(July), p. 100565.  
URL <https://doi.org/10.1016/j.asw.2021.100565>
- [183] CHEN, Y. J. I., Y. H. CHEN, J. L. ANTHONY, and N. A. ERAZO (2022) “Evaluation of the Computer-Based Orthographic Processing Assessment: An Application of Cognitive Diagnostic Modeling,” *Journal of Psychoeducational Assessment*, **40**(2), pp. 271–292.
- [184] SU, K. (2022) *Implementation of a Diagnostic Classification Model for Middle-School Physics*, phdthesis, University of North Carolina at Greensboro.
- [185] BLOOM, B. S. (1956) *A Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*, David McKay, New York.
- [186] ANDERSON, L. W. and D. R. KRATHWOHL (2001) *A taxonomy for learning, teaching, and assessing : a revision of Bloom’s taxonomy of educational objectives*.
- [187] GARFIELD, J. B., D. BEN-ZVI, B. CHANCE, E. MEDINA, C. ROSETH, and A. ZIEFFLER (2008) *Developing students’ statistical reasoning: Connecting research and teaching practice*.
- [188] MARRIOTT, J., N. DAVIES, and L. GIBSON (2009) “Teaching, learning and assessing statistical problem solving,” *Journal of Statistics Education*, **17**(1), pp. 1–18.
- [189] HADLEY WICKHAM (2016) *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
- [190] R CORE TEAM (2021), “R: A Language and Environment for Statistical Computing,” .  
URL <https://www.r-project.org/>

- [191] J, L. (2006) “Plotrix: a package in the red light district of R,” *R-News*, **6**(4), pp. 8–12.
- [192] AMERICAN EDUCATIONAL RESEARCH ASSOCIATION AMERICAN PSYCHOLOGICAL ASSOCIATION NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION and N. C. O. M. I. E. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION (2014) *Standards for Educational and Psychological Testing*, Washington, DC.
- [193] LIVINGSTON, S. A. (2004), “Equating Test Scores (Without IRT),” .
- [194] FALLSTROM, S., S. FIROUZIAN, K. KUBO, and R. PECK (2021), “Increasing Student Engagement at Two-Year Colleges Using Socially Relevant Contexts,” .
- [195] ZIMMERMAN, D. W. and R. H. WILLIAMS (1982) “Gain Scores in Research Can Be Highly Reliable,” *Journal of Educational Measurement*, **19**(2).
- [196] WILLIAMS, R. H. and D. W. ZIMMERMAN (1996) “Are simple gain scores obsolete?” *Applied Psychological Measurement*, **20**(1), pp. 59–69.
- [197] ZIMMERMAN, D. W. and R. H. WILLIAMS (1998) “Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores,” *British Journal of Mathematical and Statistical Psychology*, **51**, pp. 343–351.
- [198] GELMAN, A. and J. HILL (2006) *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press.
- [199] IMBENS, G. W. and D. B. RUBIN (2015) *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.
- [200] FORSTER, M. and C. J. WILD (2010) “Writing about findings: Integrating teaching and assessments,” in *Assessment Methods in Statistical Education: An International Perspective* (P. Bidgood, N. Hunt, and F. Jolliffe, eds.), chap. 8, Wiley, pp. 87–102.
- [201] PETOCZ, P., A. REID, and I. GAL (2018) “Statistics Education Research,” in *International Handbook of Research in Statistics Education* (D. Ben-Zvi, K. Makar, and J. Garfield, eds.), chap. 3, Springer, pp. 71–100.

## Sayali Phadke

---

CONTACT INFORMATION	Assistant Professor of Statistics, Department of Mathematics Penn State Erie - Behrend College	sayalip@psu.edu LinkedIn profile - sayali-phadke
EDUCATION	<b>Pennsylvania State University</b> , State College, PA Ph.D. Statistics with minor in Social Data analytics <b>St. Xavier's college</b> , Mumbai, India BA Economics and Statistics with Statistics Honors	<i>Defense passed:</i> August 2022 May 2011
AWARDS	2020-21 Harold F. Martin Graduate Assistant Outstanding Teaching Award Pennsylvania State University 2018 Harkness Award for Excellence in Statistics Instruction Department of Statistics, Pennsylvania State University	2021 2018
FELLOWSHIPS AND GRANTS	Seed grant, C-SoDA Accelerator Award Program Center for Social Data Analytics, Pennsylvania State University Affiliate, NSF Big Data Social Science IGERT Fellowship Pennsylvania State University Founding fellow, Young India Fellowship (Liberal Arts and Leadership) Ashoka University in collaboration with University of Pennsylvania	2021 2015 - 2017 2011 - 2012
RESEARCH EXPERIENCE	<b>Fellow:</b> Data Science for the Public Good Program, Virginia Tech <b>Research Assistant:</b> BDSS IGERT, Pennsylvania State University <b>Research Assistant:</b> Applied Statistics and Computing Lab, ISB, India	May 2017 - July 2017 Aug 2015 - May 2016 June 2012 - May 2014
TEACHING EXPERIENCE	<b>Course instructor</b> , Department of Statistics, Pennsylvania State University <b>Teaching Assistant</b> , Department of Statistics, Pennsylvania State University <b>Teaching Assistant</b> , Graduate Certificate in Business Analytics, ISB, India	Summer 2018, 19, 20 Fall 2014 - Spring 2022 Sept 2013 - April 2014
WORK EXPERIENCE	<b>Data Science Summer Associate</b> , <i>Catalist LLC, Washington DC</i> <b>Statistical Consultant</b> , <i>Society for Elimination of Rural Poverty (SERP)</i> <b>Statistical consultant</b> , <i>Ashoka Foundation: Innovators for the Public</i> <b>Statistical consultant to the Dean's Office</b> , <i>Indian School of Business</i>	May 2016 - July 2016 April 2014 - May 2014 July 2013 - Aug 2013 April 2013 - May 2014
OUTREACH & SERVICE	Co-founder, Statistics Education: Engagement and Development for Students (SEEDS) President, Society for Indian Music and Arts, Pennsylvania State University Co-founder, President, Dance Instructor <i>Nritya</i> An organization dedicated to traditional Indian dance forms, Pennsylvania State University Organizer, ASA DataFest, Hosted by Pennsylvania State University	June 2021 - Jan 2016 - August 2019 Oct 2016 - August 2019 2016 - 2018
JOURNAL PUBLICATION	<b>Phadke, S.</b> , & Desmarais, B. "Considering Network Effects in the Design and Analysis of Field Experiments on State Legislatures." <i>State Politics &amp; Policy Quarterly</i> , 19(4):451-473, 2019.	
PRESENTATIONS	[2] <b>Phadke, S.</b> , Morgan, K.L., & Hunter, D. "Causal Inference via Modeling Spillover of Treatment through Networks." North American Social Networks Conference, Washington DC [1] <b>Phadke, S.</b> , & Desmarais, B. "Network Effects in Field Experiments on Interactive Groups: Cases from Legislative Studies." Political Networks Conference & Joint Statistical Meeting	Nov 2018 2016
EDUCATIONAL MATERIAL	<b>Phadke, S.</b> , "Interactive tutorial to introduce R to introductory students". McMillan, C., <b>Phadke, S.</b> , Goist, M., & Denny, M.J., "Comparative Networks Dataset". Kancharla, M., & <b>Phadke, S.</b> , "Tutorials on Probability and Distributions".	2020 2017 2013