What do We Know about Context: An Integrated Analysis of Context Characteristics of Science Assessment Items

Dongsheng Dong

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Min Li, Chair

Chun Wang

Jimmy de la Torre

Amy Ko

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

What do We Know about Context: An Integrated

Analysis of Context Characteristics of Science Assessment Items

Dongsheng Dong

Chair of the Supervisory Committee:

Min Li

College of Education

Prior studies have widely documented the impact of characteristics of contextualized items on students' test performance. However, the role of individual context characteristics on assessment results is not fully understood. The purpose of this dissertation is twofold. The primary goal is to provide a holistic view on the influence of contextualized items on science assessment results by investigating the effects of six item characteristics in three studies. The second goal is to explore different methodological approaches to understanding effects of item characteristics and advance the psychometric analysis of contextualized items. To this end, three methodological approaches are used: item response theory, cognitive diagnostic modeling, and machine learning. Specifically, the first study addresses a common myth about contextualized

items—richer contexts are always better—by experimenting how physics items with increasing levels of richness impact students' performance. The richness of item contexts was operationalized by level of context abstractness and inclusion of illustrations. The Rasch model in combination with a hierarchical generalized linear model offers a comprehensive interpretation of the influence of context richness, such as the negative relationship between using contextualized-illustrated items and students' performance. The second study applies cognitive diagnostic models (CDM) to examine the role of context familiarity on students' mastery pattern of required physics concepts and two item parameters estimated from CDM. Results show familiarity with item context may impact students' mastery of certain physics concepts and was negatively related to the guessing parameter estimated from CDM. The third study explores the effects of three item characteristics on item difficulty based on NAEP released science items and illustrates the benefits of using cross-validation for model comparison and feature selection. The three item characteristics are cognitive demands due to the assessed topics and science practices, item format (e.g., multiple choice), and linguistic complexity (e.g., average age-of-acquisition for all words in an item). Results confirm the significant influence of item format and cognitive demands on item difficulty. Experiments were conducted to explore potential reasons why linguistic complexity was not uniquely predictive of item difficulty. Practical implications on test development are provided to serve the larger research community.

TABLE OF CONTENTS

List of Figures iii

List of Tables iv

Chapter 1 Introduction 1

Chapter 2 Understanding the Effect of Context Richness on Student Performance in Physics Items: An Item Response Approach 8

2.1 Introduction 10

2.2 Background 11

2.3 Method 23

2.4 Results 34

2.5 Discussion 44

Chapter 3 Exploring the Role of Context Familiarity on Science Assessment Outcomes: A Cognitive Diagnostic Modeling Approach 52

3.1 Introduction 54

3.2 Background 56

3.3 Method 74

3.4 Results 85

3.5 Discussion 99

Chapter 4 Analyzing Sources of Difficulty in NAEP Science Assessment Items: A Machine Learning Approach 105

4.1 Introduction 106

4.2 Background 108

4.3 Method 119

4.4 Results 127

4.5 Discussion 139

Chapter 5 Conclusion 142

References 147

Appendix A: Item Fit Statistics for Rasch Model 162

Appendix B. Results of Differential Item Functioning Analysis 163

Appendix C. Q11 in Force Concept Inventory 164

Appendix D. Selection of Linguistic Features 165

LIST OF FIGURES

Figure 2.1 Interaction between topic and context. 38

Figure 2.2 Item characteristic curves and test information curve for FM items. 40

Figure 2.3 Item characteristic curves and test information curve for E items. 41

Figure 3.1 Example of an item bundle developed based on the proximity of the items to the enacted curriculum: Close (C), Proximal (P1), and Proximal 2 (P2) (Li et al., 2012a)..... 60

Figure 3.2 Shared general information and example cases in two different contexts (Song & Bruning, 2016). 61

Figure 3.3 Prototype of a cluster (Ruiz-Primo et al., 2019). 77

Figure 3.4 The box cluster with the general context and a sub-testlet tapping acceleration 78

Figure 3.5 Attribute mastery probability for the whole student sample 89

Figure 4.1 Sample NAEP item. Content topic is earth and space science; practice is using science principles; and response type is multiple choice. 120

Figure 4.2 Decision trees learned when given different feature sets. 133

Figure 4.3 Violin plot for linear regression coefficients for 200 models trained to predict item p-values using 88 randomly sampled items for each model. 134

LIST OF TABLES

Table 2.1 Three variations of item FM42 covering the same concepts with varying degrees of contextualization 26

Table 2.2 Distractors of item FM42C with targeted misconceptions and calculated proportion of correct response 27

Table 2.3 Forces and motion (FM) items in each of the two booklets 28

Table 2.4 Energy items in each of the two booklets 28

Table 2.5 Proportion of correct response for the most attractive distractor in each item 35

Table 2.6 Coefficient estimates of HGLM 37

Table 2.7 Item parameters for the Rasch model for FM and E items 39

Table 2.8 Differences in item difficulty for FM and E items 43

Table 2.9 Correlation between average item difficulty and the difference in item difficulty between every two variants of an item 44

Table 3.1 Definitions of eight fundamental ideas 76

Table 3.2 The Q-matrix for the box cluster 80

Table 3.3 Indicators of context familiarity for each item by aggregating students' responses to survey questions 84

Table 3.4 Absolute model fit statistics for the selected models 87

Table 3.5 Final CDMs with discrimination index 88

Table 3.6 Guessing and slip parameters for all items 91

Table 3.7 Top 10 latent classes and posterior probabilities 91

Table 3.8 Classification rate for each attribute 92

Table 3.9 Results of regressions using students' survey responses to predict probability of mastering A1 to A8 96

Table 3.10 Results of linear regressions using two aggregated familiarity indicators to predict guessing parameter and slip parameter 98

Table 3.11 Statistical summary of three evaluation criteria for parameter recovery in the simulation study 98

Table 4.1 Existing studies on aggregated item difficulty prediction, listing number and type of items, predicted variable, and predictors used in the study 114

Table 4.2 Existing studies on predicting individual student performance 115

Table 4.3 NAEP data—132 items broken down by content, practice and response type 120

Table 4.4 R^2 for different prediction models and feature sets, comparing results when fitting to the full data set vs. 3-fold CV, showing that the training R^2 is not a reliable indicator of test R^2 128

Table 4.5 R^2 for decision tree with and without depth selection using all features 128

Table 4.6 Metrics for all regression models 130

Table 4.7 Multiple linear regression analysis results for models trained with format and cognitive features only (left), these features plus the AoA linguistic feature (center), and all linguistic features (right) 131

Table 4.8 Model selection criteria for LR and DTs with different linguistic features combined with format and cognitive features 132

Table 4.9 Correlation values of linguistic features with linear regression coefficients of format and cognitive features for 200 linear regression models each trained on 88 randomly sampled items 136

Table 4.10 NAEP and PISA item difficulty prediction with and without the age-of-acquisition (AoA) linguistic feature 139

ACKNOWLEDGEMENTS

I own a debt of gratitude to many people, who inspired, supported, and accompanied me throughout my PhD journey. First and foremost, I would like to thank my committee: Min Li, Chun Wang, Jimmy de la Torre, and Amy Ko. This dissertation would not have been possible without your insightful suggestions and generous support. A special thanks to Dr. Chun Wang for providing valuable feedback and guiding me through the completion of Chapter 3. I am also grateful to Dr. Wenchao Ma, whose expertise and knowledge was a tremendous help in this dissertation process.

I am deeply indebted to my advisor, Dr. Min Li, for her continuous support during my academic study and enormous contribution to this dissertation. She sees the potential in me that I did not see and makes me believe in it. She always encourages me to explore new research areas and challenges me to think deep about my research. Her knowledge and guidance are essential to the completion of this dissertation and have taught me valuable lessons about life and research. I appreciate all the conversations we had, all the projects we collaborated, and all the candies we shared during meetings. I feel so fortunate to be her student and to have learned from her passion and work ethnics.

My sincerest gratitude goes to the DECISA project for laying the foundation for my dissertation and providing data for the first two studies. Throughout the project I have met excellent mentors and colleagues including Maria Araceli Ruiz-Primo, Jim Minstrell, Xiaoming Zhai, Klint Kanopka and Phillip Hernandez. They set a high standard for research and teach me the importance of learning from different perspectives. I would also like to extend my thanks to

Dr. Mari Ostendorf and Farah Nadeem, who made tremendous contributions to Chapter 4 and opened a door for me to venture into the research area of machine learning.

My graduate life at University of Washington would have not be so memorable and cheerful if without my cohort and my friends. Thanks to Phonraphee Thummaphan, Linda Liaw, Ting Wang, Linghui Zhu, Weijia Wang, Soo-Yeah Shim, Jiyoung Lee, Nathan Abe, Gabby Gorsky, Nixi Wang, Wei Zuo and many other friends who shared laughs and tears with me and supported me in all possible ways. I would not have come this far without the cheers and supports provided by two friends in particular: Alec Kennedy and Cricket Limlingan. They listen to my needs, understand my struggles, and always encourage me to stay positive. I cherish all the time I had with these lovely people.

Finally, I own indefinite gratitude to my family. Thanks to my mom, my dad, and my brother for respecting every decision I made and backing me up whenever I need help. My deep gratitude goes to my husband, Chaoyi Huang, who believes in me when I doubt myself, lifts me up when I feel frustrated, and supports me with love and understanding. Finally, a special thank you to my little one, Anrui Huang. You are the biggest motivation of this dissertation. Because of you, I hope I could make a little difference to the current assessment system so that kids like you could receive fair and better education and be recognized as who you are.

DEDICATION

To my son

Chapter 1
Introduction

Standardized assessments with abstract items used to be widely adopted to assess students' achievements and compare the adequate yearly progress across states and schools (Garcia, 2014). Although it offers an objective measure of students' learning and is easy to interpret, it is criticized for assessing limited aspects of students' learning (Garcia, 2014) and disconnecting learning from real life. As Wiggins (1993) points out, such traditional tests are more like drill tests that do not prepare students for real, "messy" application of knowledge in contexts. Indeed, the results of standardized assessments with abstract items cannot tell us how students would use the knowledge to solve a problem in real-life situations, as problems in the real world never appear in the form of an abstract formula or an abstract situation (for example, students are asked to ignore friction between object A and the ground). In response to the criticism on the traditional test design, research (e.g., Almuna Salgado, 2017) and educational policies (e.g., National Science Education Standards, 1996) advocate for the incorporation of authentic contexts in assessment items.

A *context* refers to a set of characteristics of an item that provide supplemental information to frame the question and response choices (Kirsh, 2009; Wang, 2016). In educational assessments, item context usually defines a scenario or provides a description of a practical problem, a natural phenomenon, or a lab setup (Ruiz-Primo & Li, 2012, 2015, 2016). Item with contexts have been referred under different names: *contextualized* (e.g., Ruiz-Primo & Li, 2016), *scenario-based* (e.g., Mcmartin, Mckenna, & Youssefi, 2000), or *story problems* (e.g.,

Caldwell & Goldin, 1987). In this dissertation, I follow the definition proposed by Ruiz-Primo and Li (2012, 2015, 2016) and refer such items as *contextualized items*.

The use of contextualized items has been widely supported in the literature. Based on cognitive psychology theories, Klassen (2006) claims context and cognition should not be separated as cognitions are contextual dependent. "It is possible that multiple representations of a concept exist in different context or mental networks; for example, scientific concepts may exist both in its outside-of-school experiential form and in its school-science form" (Klassen, 2006, p. 832). Empirical studies further support that a good contextualized item scaffolds the problem-solving process in an assessment by reducing cognitive load, facilitating semantic comprehension and allowing for the construction of mental representations that aids problem solving (Cook, 2006; Mevarech & Stern, 1997). From a test development perspective, contextualized items are argued to improve the validity of the test. Wiggins (1993) notes the conventional test design has a tendency to sacrifice validity for reliability. The test items are designed to be generic and with little ambiguity as possible to ensure the precision of scores, whereas such a test form fails to provide information about students' true state of acquisition of the assessed knowledge. Contextualized items address this issue by offering an opportunity to measure students' ability to transfer what they have learned to a variety of contexts. Such transfer of learning across contexts serves as an essential indicator of the depth of students' understanding. Lastly, I argue that using contextualized items could potentially promote learning during the assessment. As contextualized items open up the potential to link school learning with real-life situations, students' understanding of the knowledge may be strengthened by connecting to and reasoning through the contexts.

Despite the wide use of contextualized items in large-scale assessments such as National Assessment of Educational Progress (NAEP) and Programme for International Student Assessment (PISA), there is an increasing concern that the current findings on the effects of contextualized items may not capture the whole picture. Extensive studies have shed light on a variety of factors that have an impact on students' performance. Some factors that are widely studied are the use of external representations (e.g., Berends & van Lieshout, 2009), familiarity with the context (e.g., Boaler, 1993), linguistic load embedded in the text (e.g., Abedi & Lord, 2001), and item format (Mesic & Muratovic, 2011), etc. All these factors are essential *context characteristics* of a contextualized item, which we will call *item characteristics* for brevity. Every single characteristic plays a role in defining the context of the item and deciding how the item is perceived by an individual student. Such a connection between item characteristics and students is important yet hard to detect. In fact, if we trace back to the previous literature on contextualized items, we may find that most studies on item characteristics report mix findings in terms of their influences on assessment outcomes. For example, illustrations are found to engage students in the test and motivate students to try harder to comprehend the text (Brookshire, Scharff & Moses, 2002; Cook, 2006), while illustrations with unnecessary details may drive students' attention away from the intended constructs (Ahmed & Pollitt, 2000). As Ruiz-Primo and her co-workers (Ruiz-Primo, Li, Minstrell, Dong, Kanopka, Hernandez, & Zhai, 2019a) point out, "context is a delicate component of items as it may affect students' performance in a positive or in a negative manner, by introducing either construct-relevant or irrelevant variance, respectively" (p. 3).

Current research on item characteristics is rich but scattered. Although hundreds of studies have focused on different kinds of item characteristics, few of them provide a holistic

view regarding the influence of contextualized items by taking into count the investigations of multiple item characteristics. While learning about individual item characteristic is critical, a holistic understanding of contextualized items offers important implications for item writing. When it is impossible to take into account the effects of all item characteristics, having a holistic understanding about impacts of contextualized items allows item writers to focus on the big picture, paying more attention to item characteristics that are critical in that context while considering potential risks posed by other characteristics. This dissertation demonstrates an effort to provide a holistic interpretation of how contextualized items impact assessment results. As it contains three studies which explored different item characteristics with different experimental designs, it connects three studies loosely by framing them under a universal definition of context. Admittedly such a loose connection does not reflect academic rigor, I hope it brings inspirations to the future work in terms how to relate results of multiple studies.

Regardless of how much work has been done on item characteristics, we have to admit that detecting the effect of an item characteristic on assessment results is difficult. The difficulty comes from the fact that its impact varies by content topics, by student differences, and by many other factors such as how it interacts with other item characteristics (e.g., Dong, Li, Ruiz-Primo, Zhai, & Minstrell, 2018; Almuna Salgado, 2017). Such a challenge inspires researchers to examine the effects of item characteristics from various perspectives and develop innovative methodologies to uncover the true story. Prior work on contextualized items has concentrated on two major outcomes of assessments. One is directly related to students' individual performance, usually indexed by students' scores or the probability of students responding to an item correctly. The other outcome deals with item-level analysis statistics—the most widely reported one is the item difficulty index. Students' performance and item statistics are two important aspects of

assessment results. They bring different implications for test interpretation and test uses (Kane, 2006): student performance allows us to infer about students' ability and develop tailored instructions for a particular or even larger student group; in contrast, item statistics reflect the appropriateness of assessment items pertaining to the general student population, which further informs the test development and validation.

Targeting different outcomes motivates the use of different methodologies. To predict students' performance, item characteristics are usually incorporated as a variable to predict student responses to assessment items using models such as regressions, analysis of variance (ANOVA) or *t*-tests. For instance, Milenković, Segedinac, Hrin and Gajić (2016) performed the Kruskal-Wallis ANOVA to explore the role of context on perceived cognitive load and students' performance in problem-solving tasks. Tasks were classified into three groups based on three levels of complexity in context: without context, with moderate context, and with rich context. To analyze students' achievements, the ANOVA was conducted first to determine whether there was a significant difference between three groups of tasks in terms of students' performance. Then post-hoc pairwise comparisons were performed to detect which group of tasks were significantly different from others. With regards to item statistics, the impact of item characteristics on item statistics is usually analyzed using the Item Response Theory (IRT) framework. One of the most commonly used models is Fischer's (1993) linear logistic test model (LLTM). Chen, MacDonald, and Leu (2011) applied the LLTM to investigate the source of item difficulty in mathematical fraction items and found six cognitive operations significantly impacted item difficulty. Six cognitive operations were: using illustrations, providing interpretations, applying judgment, computation, checking distractors, and solving routine problems. In another study, Baldwin (2008) applied a different IRT model (two-parameter

Bayesian Testlet model) to explore the effects of three item characteristics on item statistics: vignette word count, stem word count, and options word count. This model not only explained item characteristic effects but also accounted for the fact that multiple items shared a same context. It should be noted Baldwin (2008) looked into three item statistics—discrimination, difficulty and the testlet effect parameters—rather than solely focusing on the item difficulty.

This dissertation aims at providing empirical evidence to contribute to the current research conversation on contextualized items. The purpose of this dissertation is twofold. The primary goal of this dissertation is to examine the effects of six item characteristics on students' performance and/or item statistics using science assessment items. The six item characteristics were explored through three studies: the first study focuses on *the level of abstractness* and *use of visual representations* of contextualized items; the second study examines the role of *context familiarity* on two assessment outcomes; and the third study investigates the effects of three item characteristics—*cognitive demands, item format, and linguistic complexity*—on item difficulty. Furthermore, the second purpose of this dissertation is to demonstrate different methodologies to predict students' performance and/or item statistics and make inference about effects of item characteristics. The data analysis methodologies include statistical models commonly used in educational measurement (e.g., Rasch model) and also other innovative approaches such as the cognitive diagnostic modeling and cross validation. All models and analysis procedures could be grouped under three general frameworks—item response theory, cognitive diagnostic modeling, and machine learning. I believe it is important to illustrate different data analysis approaches because it opens up the possibilities to understand effects of item characteristics on learning outcomes other than student scores and item difficulty. For example, in this dissertation I looked at the effects of item characteristics on students' mastery profile of assessed physics concepts

and item parameters such as how likely students could guess an item correctly in general. Moreover, using different approaches to tackle the same problem helps validate the results and provide confidence for the interpretation of the findings. Chapters 2 to 4 demonstrate the application of the three frameworks accordingly.

Chapter 2

Understanding the Effect of Context Richness on Student Performance in Physics Items: An Item Response Approach

Dongsheng Dong[1], Min Li[1], Maria Araceli Ruiz-Primo[2], Xiaoming Zhai[3], & Jim Minstrell[4]

[1] College of Education, University of Washington

[2] Graduate School of Education, Standard University

[3]Department of Mathematics and Science Education, University of Georgia

[4] Facet Innovations

Abstract

In educational assessments, item context usually defines a scenario or provides a description of a practical problem, a natural phenomenon, or a lab setup. This paper examines how levels of context richness influence students' test performance on a physics exam for 383 7th grade students. Three item contexts were defined based on increasing richness: The Abstract (A), Contextualized (C). and Contextualized and Illustrated (CI). Effects of item contexts were analyzed at the overall test level and at the item level. Test-level results indicate that using contextualized items (C) improved test performance compared to using abstract items (A), while the contextualized and illustrated items (CI) reported poorest performance among three types of items. Moreover, a joint effect of contexts and topics was detected. Using contextualized items led to a better performance regardless of topic, but the magnitude of increase in performance was greater under the Energy topic than under the Forces and Motion topic. Similarly, adding illustration to contextualized items makes items more difficult under Energy topic than under the

Forces and Motion topic. The item-level analysis further suggests that the effect of contexts vary by items, which could be due to factors such as content difficulty (cognitive load due to intended science concepts), complexity of contexts (combinations of item characteristics), and learner differences.

2.1 INTRODUCTION

Contextualized items are widely used in assessments. According to Wang and Li (2014), about 70% of Grade 4 and 8 science items released by National Assessment of Educational Progress (NAEP) are contextualized. Similarly, more than 70% of Grade 8 science items released by Trends in International Mathematics and Science Study (TIMSS) in 2011 are situated in contexts. The increasing popularity of contextualized assessment items is based on the premise that it facilitates learning and provides a comprehensive evaluation of students' true ability. For instance, prior work argues embedding assessment items in contexts aids comprehension and problem solving (e.g., Brookshire, Scharff & Moses, 2002), reduces cognitive demands of assessment items (e.g., Ruiz-Primo & Li, 2016), and allows for the measure of transfer of learning (Wang, 2016). However, as Ahmed and Pollitt (2007) point out, over-advocating the advantages of contexts may be dangerous as we may forget the fact that context itself adds extra demands.

In this study, we would like to draw the attention back to some fundamental questions: does using contextualized items really brings more benefits than using abstract items? If yes, then is it the case that the richer the context, the more benefits it brings to students? The purpose of the present study is to explore how the level of richness of item contexts influence students' test performance on a middle-school physics exam.

We follow Ruiz-Primo and Li's (2015, 2016) framework to define context. Their framework lays out three general features of contextualized items: *level of abstractness* (whether abstract ideas are present and whether the abstract ideas identified have any concrete reference such as objects or events), *type of resources* (whether the item includes any nonverbal

representations such as table, graphs, etc.) and *nature of the context* (how the setting is described in the context: everyday house activity, professional/workplace or scientific information, etc.).

Following Ruiz-Primo and Li's framework (2016), specifically *level of abstractness* and *type of resources*, we define three levels of item contexts based on increasing richness: Abstract (A), Contextualized (C), and Contextualized and Illustrated (CI). *Abstract* items express ideas without any scenario-based information or concrete reference such as objects and events, and thereby are considered the least contextually rich. *Contextualized* items situate problems in real-world settings (e.g., everyday activity, lab settings, etc.) and usually are accompanied with a concrete description of the background information and/or a storyline. Lastly, *contextualized-illustrated* items add another layer of richness to contextualized items by providing illustrations to guide students' visualization, which are considered richest in context. The two guiding research questions are:

(1) Does adding contexts to items improve students' test performance compared to abstract items?

(2) How does the integration of illustrations further influence students' performance compared to students' performance on abstract items and contextualized items?

2.2 BACKGROUND

This section summarizes the findings of prior work on the use of contexts in assessment items and the integration of visual representations to contextualized items. For each item characteristic, benefits and potential challenges are discussed based on the evidence collected from empirical studies.

2.2.1 Abstract versus Contextualized

Why Contextualized Assessment Items? The advantages of adding contexts to assessment items can be summarized into three aspects. First, using contextualized items increases test motivation. Loaded with isolated facts and abstract concepts, science subjects may be perceived difficult, irrelevant or even boring for students (Barmby, Kind & Jones, 2008; Vos, 2014). According to Vos (2014), there is a trend that students' attitudes towards science decline throughout their progression of science learning. Such a change in attitudes may be partly attributed to the disconnection between abstract concepts learned in school and students' everyday life. The consequence of low motivation causes concerns on the validity of test scores: students tend to make less efforts to respond to test items, leading to an inaccurate measure of their content knowledge and learning ability. The introduction of contextualized items is intended to address this concern by making test items more relevant to students' everyday life. Students are motivated and kept engaged as they perceive such contexts as more realistic, interesting, and useful than typical abstract textbook problems (Haladyna, 1997). A study by Nijlen and Janssen (2015) provides evidence that items situated in authentic problem-solving contexts could reduce the construct-irrelevant variance caused by low test motivation. The results show that overall students reported a lower item omission rate on contextualized items than non-contextualized items. Nijlen and Janssen (2015) concluded that contextualized items might be less prone to low examine effort in low-stake testing situations, as "even those students that consider themselves to be not very good at mathematics seem to make an effort on the contextualized items, while for the non-contextualized items they do not even bother to start working on a lot of the items" (p. 82).

Second, it is argued that contextualized items measure students' ability to apply knowledge to various situations (Ahmed & Pollitt, 2000; Nijlen & Janssen, 2015). According to Ahmed and Pollitt (2007), abstract questions can often be answered by directly repeating passages from the textbooks. A test consisting of abstract items may not be able to reflect students' true ability as it becomes "a catechism of question and learned answers" (p. 202). Situating items in contexts could minimize the risks of testing rote memory and offer opportunities to assess deep learning inherent in transfer of knowledge (Wang, 2016). As an example, Heller and Hollabaugh (1992) examined how contextualized items facilitated problem solving by observing 400 college students from two physics classrooms of two schools working in small groups. Students were asked to collaborate to solve two types of physics questions: traditional abstract problems and contextualized problems. It was found that the contextualized group demonstrated a better use of problem-solving strategies than the abstract group. Specifically, students working with contextualized problems spent most of time discussing the physics concepts and principles that were needed to solve the question, whereas the discussion of students working with traditional physics problems mainly focused on what formulas to apply. As Wiggins (1993) points out, "we cannot be said to understand something unless we can employ our knowledge wisely, fluently, flexibly, and aptly in particular and diverse contexts" (p. 200). Compared to abstract assessment items emphasizing the application of formations and equations, contextualized items gauge the depth of conceptual understanding by linking the abstract knowledge to the event and encouraging students to employ different cognitive skills related to the target concepts to solve the problem.

Third, contextualized items reduce cognitive load. Caldwell and Goldin (1987) examined the effect of item context on students' test performance on four types of items: abstract factual,

abstract hypothetical, concrete factual and concrete hypothetical. An item is considered factual if it merely describes a situation. A hypothetical item not only describes the situation, but also indicate a possible change in the situation that does not really occur within the context of a problem. The results indicated that overall students performed better on concrete items compared to abstract questions. Moreover, it was found students tended to solve more concrete factual items correctly than concrete hypothetical items. This brings in the implication that contextualized items are particularly helpful when it mimics real-world situations as it evokes students' relevant knowledge and real-world experiences, which in turn enhances the comprehension of what the item asks. Consistent with the study of Caldwell and Goldin (1987), Ruiz-Primo and Li (2016) investigated 52 PISA science items released from 2006 to 2009 and examined how context characteristics is associated with students' test performance and cognitive demands. They found that concrete ideas were associated with lower cognitive load, but no further explanations were provided for the underlying mechanism behind the association. Klassen (2006) states that contexts allows for the connection and integration of new information to similar information stored in someone's long-term memory. The benefit of such integration is that it enables the retention of information and enables students to use prior experiences for meaning making and comprehension. This is confirmed by Ahmed and Pollitt (2007), who found items with focused contexts could activate relevant concepts and scaffold anticipated cognitive processes.

Issues with Contextualized Assessment Items. Although contextualized items are believed to aid comprehension by making reference to the real world, the potential drawbacks should not be neglected. A number of researchers doubt whether it really serves the intended purpose in assessments. Boaler (1994) claims that contextualized items in school mathematics is

set up in a way that it presents real-life variables but do not require students to take them into consideration when solving the problem, which interferes with their common-sense beliefs. In some cases, some "real-world" contextualized items may require students to "suspend reality and ignore their common sense in order to get a correct answer" (p. 554). She offered a mathematics problem as an example: *The cold tap on full will fill a bath in 5 minutes. The hot tap takes 20 minutes. When the taps are off, a full bath takes 8 minutes to empty. If both taps are turned on full but you forget to put the plug in, how long will it be before the bath overflows?* The common-sense answer for this question was that the bath will never overflow if the plug was not put in. However, she commented that in assessments this was not even the right mode of thinking when approaching this question. Boaler (1994) thereby further noted that when solving math word problems students tended to be involved with the real-world context and take into account real-world variables while they were not supposed to, which may lead to a lower performance on contextualized items in contrast to abstract items.

A related criticism opposing the use of contextualized items is that adding contexts to questions may distract students from the scientific ideas intended by the question and focus on the context rather than the content. Mevarech and Stern (1997) examined how sparse and real contexts affect students' understanding of abstract mathematical concepts. They implemented four experiments: two with 12-year-old children and two with undergraduate students. The participants were asked to interpret isomorphic linear graphs in sparse and real contexts. With the intention to examine context effect on different age groups, results of all four experiments showed that tasks embedded in sparse contexts were easier than the tasks in real contexts, as both age groups reported higher performance on sparse contexts. Mevarech and Stern (1997) posited three explanations for the difficulty associated with real contexts. First, students' attention was

driven away from the mathematical features of the problem as we discussed above. Second, the real context might activate a simplistic mathematical model with which students focused on isolated elements in the item rather than considering the task as a whole. More importantly, it was found that different contexts triggered different knowledge structures for both age groups. The sparse context was more associated with logic-mathematical explanation while the real context led students to base their explanations on logical inference, practical justifications, or reference to specific points.

In line with the findings by Mevarech and Stern (1997), Ahmed and Pollitt (2007) point out that contextualized items may contain irrelevant information that confuse and mislead students. The negative consequence associated with irrelevant contexts may be particularly salient for young learners. As children tend to have difficulty suppressing irrelevant information in the text, their attention may be easily driven away from the underlying concepts being tested in the question, thus hindering their comprehension and test performance (Berends & van Lieshout, 2009).

Finally, while contextualized items increase the concreteness of the content, it also increases the linguistic demand. The reading load comes from the fact that the text not only becomes lengthier, but also more complex, as the languages used in a real-world context may involve metaphors or culturally specific terms (Ahmed & Pollitt, 2007). The great demand on verbal skills from contextualized items may disadvantage students with low linguistic proficiency, particularly ELLs, and result in negative impacts on test motivation (Nijlen & Janssen, 2015). This alarms test writers to be cautious about the linguistic load when constructing contextualized items. When students do not possess the necessary verbal skills to

decode the context, the context may become inaccessible to students, making the item unnecessarily difficult.

2.2.2 Illustrations as a Contextual Feature

Benefits of Illustrations. Common visual representations used in science instructions and assessments include pictures, illustrations, graphs, symbols, tables and diagrams (Ruiz-Primo & Li, 2012, 2015; Wang, 2016). As multimedia learning is largely incorporated in science classrooms, the combination of verbal and visual representations is considered an essential approach to better communicate science concepts and scaffold students' scientific thinking. In science education, visuals like illustrations are argued to be effective ways to represent scientific phenomena that are hard to be observed or experienced directly and to model abstract relationships and processes that are difficult to conceptualize (Cook, 2006).

The inspiration of using visual representations as learning aids comes from the dual coding theory (Paivio, 1990; Mayer, Sims, & Levin, 1994). According to this theory, verbal and visual information are processed cognitively by two different subsystems: verbal and imagery systems. The separate coding systems lead to the construction of different mental models: learners build a verbal representational connection by transforming texts into a mental representation of the verbal materials; similarly, presented visual information are internalized by building a visual representational connection. Meaningful learning occurs when referential connections are constructed between the verbal and visual representations. As visual representations gain increasing popularity in instructional practices, researchers extend its use to assessment purposes. Two theoretical explanations that support the use of visual representations in assessments are motivation hypothesis and the repetition theory. The motivation hypothesis claims that the use of illustrations may improve test performance as students may be more

engaged in the test and are motivated to try harder to understand the text (Brookshire et al., 2002; Cook, 2006). The repetition theory believes the presentation of both texts and visuals double the exposures of the information to students, therefore enhancing memory (Brookshire et al., 2002). These theories explain the underlying mechanism of how visual representations impact learning and cognition. Based on these theories, incorporating illustrations in assessment items may assist in comprehension and reducing cognitive load by providing an alternative channel to decode information other than text.

A study by Brookshire et al. (2002) suggested that books with texts and illustrations improved reading comprehension for young learners. In the study, 71 first and third graders were presented with one of nine books varying in types of representations (text-only, illustration-only, and text-plus-illustration) and types of illustrations (varying in brightness and styles). Students were asked 15 comprehension questions after reading (third graders) or being read to (first graders) the book. The 15 comprehension questions consisted of five questions per type of representation (questions designed for text-only, illustration-only, and text-plus-illustration conditions, respectively). Results showed that overall both first and third graders reported highest comprehension in the illustration-plus-text content compared to other two types of representations.

As the study by Brookshire et al. (2002) points out the positive impact of adding illustrations to texts on comprehension, it remains unclear what cognitive domains that illustrations contribute to. Lewalter (2003) extends previous studies by comparing the impacts of including static and dynamic visuals in an expository test on students' learning outcomes and strategy use. A total of 60 education and psychology undergraduate students participated in the study. Students were assigned to study with a computer-based learning program on an

astrophysical subject with one of the three versions of the learning materials: text-only version, text with static illustration and text with dynamic visuals. The dynamic visuals accompanying with the learning text consists of animated graphics covering the complete course of motion; while the static illustrations describe or explain motion through a single frame or a series of frames or symbolize the development of the light ray by arrows. Students were tested on two types of tasks: (1) factual knowledge, and (2) comprehension and problem solving. According to the results, students assigned to the illustrated learning materials reported significantly better performance than the text-only group on factual knowledge. Moreover, students learning with illustrations reported a high frequency of using the rehearsal strategy when working on the visuals, which refers to the strategy of "repetitions of the learning text with exact wording or the recapitulation of the main idea of the learning content while using the expressions of the text" (Lewalter, 2003, p. 179).

Studies by Brookshire et al. (2002) and Lewalter (2003) contribute two important findings to the research of illustrations. First, they confirm the potential benefits of illustrations as a contextual aid for information retrieval for both young and adult learners. Second, the fact that students frequently use rehearsal strategies when working with illustrations provide evidence to substantiate the dual-coding theory, as students seem to make referential connections between visual and verbal representations by mapping what they see in illustrations to what they are reading in texts. Their findings further support the argument that incorporating illustrations to assessment items may improves students' comprehension of the test items and reduces the risks of misunderstanding or misinterpretation.

Beyond the universal benefits of visuals to the general student population, literature also highlights its unique function in supporting certain subgroups such as English language learners

(ELLs) in assessments. Martiniello (2009) reported an experiment regarding the effect of linguistic complexity and nonlinguistic representations on differential item functioning (DIF) measures based on item response theory difficulty parameters. The study analyzed responses from 68,839 fourth graders in a state mathematics test. Results suggested a positive association between nonmathematical linguistic complexity and differences in difficulty parameter estimates favoring non-ELLs over ELLs. However, the inclusion of nonlinguistic schematic representations tended to mitigate the negative effects of increased linguistic demand for ELLs. Aligning with findings by Martiniello (2009), Solano-Flores et al. (2014) proposed to use vignette illustrations as a testing accommodation for ELLs. Crisp and Sweiry (2006) noted that students with low reading skills tended to view illustration as a helpful tool and relied more on them in search for clues than students, while students with high reading skills were more likely to think that illustrations are unnecessary.

Challenges with Using Illustrations. As the dual coding theory argues that adding illustrations is generally beneficial for learning (Schnotz & Bannert, 2003), it is criticized for overlooking the fact that "subject matter can be visualized in different ways and that the form of visualization affects the structure of the mental representation" (Schnotz & Bannert, 2003, p.153). Indeed, the complexity of illustrations has posed a great challenge to test developers and item writers. Wang (2012) examined illustrations used in assessment items from TIMSS, PISA and Chinese large-scale assessments and reported an average use of 22 features per test item. The features ranges from representation of objects and backgrounds, such as image concreteness (e.g., realistic line drawing), dimensions (e.g., three dimensional), relative scale of objects (e.g., proportionate) to context in illustration such as socio-historical context (e.g., event in domestic affairs) and so on. An overcomplicated illustration may pose unnecessary cognitive loads on

students. Solano-Flores and Wang (2015) reported a secondary analysis study which explored the effect of illustration characteristics on students' performance in PISA 2009 science assessment. Responses of 11,662 to 11,746 students, respectively from Shanghai-China, the U.S., and Mexico, were analyzed. Correlations were conducted between item difficulty and complexity of illustration. The illustration complexity was measured by about 100 types of pre-defined illustration features. According to the findings, an increase in illustration complexity was associated with greater item difficulty for U.S. and Mexico.

Based on findings of previous studies, selecting appropriate features of illustrations is important as it impacts the quality of test items and influences how students perceive the items. For example, Ahmed and Pollitt (2000) argued that inappropriate information in illustrations could mislead students by activating incorrect schema. In one of their studies, students were presented with the following item: *Thousands of golf balls end up in the sea. They are lost by golfers on seaside golf courses and on cruise ships. Fish sometimes swallow these golf balls. Explain why these could harm the fish*. Along with the item, a cartoon was presented with three fishes surrounding the golf balls and holding forks and knives as if they are going to eat the golf balls. Ahmed and Pollitt (2000) commented the personified cartoon tended to confuse students; instead of a fish schema being activated, students tended to construct a schema in which fish was humanized, which caused them to write about choking.

As another example of how visual features impact the perception of illustrations, Crisp and Sweiry (2006) investigated whether subtle changes in physical features of illustrations could influence students' comprehension of test items. A total of 525 16-year-old students responded to a test of six items involving graphical elements, of which 27 pairs of students were interviewed after taking the test. For most of these questions, two versions were constructed with subtle

differences on the physical features of illustrations. Changes included the position of a certain sentences (far above the table versus close above the table), relative size of objects (intentionally make an object look bigger than others), and presence or absence of an example illustration, etc. According to the results, most of changes did not significantly impact students' performance except for the one item. For that item, it is reported that students performed significantly better when the illustration corresponds to the real status of the object in daily life.

Beyond the complexity of illustration, an additional challenge regarding illustration comes from its interaction with test-takers and other elements of the test items. Schnotz and Bannert (2003) claimed illustration enhance learning when students have low prior knowledge and the subject matter is visualized in a task-appropriate way. Similar findings were reported in the study by Mayer (2003), which pinpointed four important conditions where illustrations facilitate learning. According to Mayer (2003), students learn more deeply when (1) verbal and visual representations are presented together rather than words alone (multimedia effect) (2) extraneous material is excluded rather than included (coherence effect), (3) textual information are positioned close to rather than far from corresponding pictures (spatial contiguity effect), and (4) words are presented in conversational rather than formal style (personalization effect). With all those findings, further investigations are required to provide empirical evidence for the effect of illustrations on students' performance in various situations and for different groups of learners.

As a summary of the discussion above, it seems difficult and inappropriate to reach any conclusions about how contexts (including verbal and visual representations) should be used and interpreted to understand students' thinking. As Booth and Koedinger (2012) point out that the benefit of visual representations "is dependent on their relevance for the task at hand, the context

of the representation, and the skills of the user" (p. 492), we believe this argument is applicable to the general discussion of context effects. The effect of item contexts on students' performance has to be fully understood in specific conditions by taking into account features of items and test-takers (e.g., content of the task, nature of the context, use of presentations, and the student group, etc.) before being generalized, which brings implications for an item-level approach to analyze context effects (Boonen et al., 2014) in addition to the test-level approach commonly used in the literature.

2.3 METHOD

2.3.1 Participants

Three hundred and sixteen students from 7th grade students, corresponding to 4 teachers, from the Seattle School District participated in this study. Each student corresponded to one period of a teacher's daily schedule. All students were confirmed that they had learned the concepts of "Forces and Motion" and "Energy" by the time they were tested. Besides students' responses to test questions, we also obtained students' demographic information on gender, whether they are enrolled in special education (SPED), and whether they are English Language Learners (ELLs). Unfortunately, the demographic information about the students was not provided by one of the participating teachers, resulting in a sample of 239 students whose demographic information is available. Among those students, 53% are female (versus 47% male), 8% are ELLs (versus 92% non-ELLs) and 13% received special education (versus 87% No-SPED).

2.3.2 Instrument

The instrument consists of eight Forces and Motion (FM) items and four Energy (E) assessment items developed by our research team with each item having three variants that reflect the three levels of context richness; in other words, each item has A, C, and CI versions, totaling 36 items. It intends to serve three important purposes. First, it provides an evaluation tool to measure students' overall understanding of two target topic domains—Forces and Motion (FM) and Energy (E). Second, it diagnoses students' misconceptions in FM and E with an intention to provide pedagogical implications for classroom learning. Third, with different manipulations of item characteristics, it intends to identify the source of confusion and difficulty embedded in item contexts and provide instructions for item writing.

Item Development. The development of items went through a rigorous iterative process of revision. We collected 5,979 items focusing on secondary science and low undergraduate level of physics from different sources (e.g., released international and national tests, website with assessment resources, and research papers). From the 5,979 items only 153 items passed multiple screenings based on its focus on the two target topics, the alignment with the NGSS standards at the middle school level, and the appropriateness of contexts. All items were directly reviewed and discussed by three expert panels which consisted of very experienced local high school teachers and physics education researchers. Based on the feedback from panel reviews, we developed eight FM items and four E items by revising existing items or constructing new items based on target fundamental physical concepts extracted from the panel review.

Context Manipulation. Each new item is manipulated into three versions by adding contexts and illustrations (e.g., FM11A, FM11C, FM11CI). The three manipulated versions assess the same fundamental science concepts but differ in degrees of context richness (A, C,

CI). As an example, Table 2.1 presents the three different richness levels of an FM item (FM42) assessed in the study. All sets of developed items were iteratively revised by experienced physics researchers, physics teacher and linguists. The development of illustrations was consulted with physics teachers and researchers specialized in visual representations.

To capture students' alternative thinking, we identified a set of common misconceptions in FM and E and weaved them into distractors. Table 2.2 shows the misconceptions targeted by FM42. Taking FM42C as an example, FM42C intends to measure students' understanding about forces involved in interactions. Distractors are based on students' common misconceptions and could be mapped back to one of the questions (Q11) in Force Concept Inventory (FCI; Hestenes et al., 1992). We calculated the proportion of correct response for each distractor in FM42C and compared it with the average proportion of correct response for each option reported in FCI. The two statistics seem to be consistent and closely matched. It shows almost half of students hold a strong misconception that the magnitude of force is associated with the mass and the motion of the object. A fair number of students (approximately 30%) believe that only active agents could exert forces. Only a small proportion of students (approximately 12%) answered this item correctly.

Table 2.1

Three variations of item FM42 covering the same concepts with varying degrees of contextualization

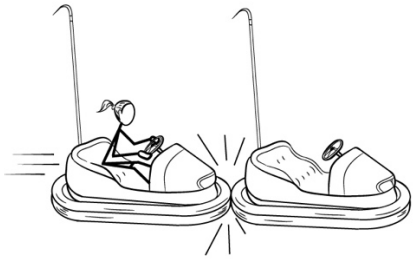
FM42A: Abstract (A)	
In the middle of a flat surface, a heavy moving object collides with light object that was not moving. The heavier object now moves slowly, while the lighter object moves fast across the frictionless level surface. Consider the forces at the moment of the collision. Which statement is correct?	*Type of Abstractness*: Ideas are expressed in abstract science languages. Objects and events are defined by its most salient physics features with no concrete references. For example, we could know the weight (e.g., "heavy" and "lighter") and the motion ("moving") of the objects but could not tell what the object is. *Nature of Context*: No information implies the setting of the event. The event is described as a scientific phenomenon rather than concrete daily activities. *Type of Resources*: No illustrations or any other visual aid.
FM42C: Contextualized (C)	
At the amusement park, Jane drives her bumper car into an empty, stopped bumper car. Jane's car now moves slowly, but the empty car moves fast across the smooth, frictionless floor. Consider the force(s) at the moment when Jane's car hits the empty car. Which statement is correct?	*Type of Abstractness*: Descriptive languages are used with concrete references to where, who and what happened, which pinpoints a complete story of the bumper car game. For example, compared to the "heavy" and "lighter" moving objects in the abstract item, we know the objects are bumper cars. The "heavy" car is driven by Jane and the "lighter" car is an empty, stopped bumper car. *Nature of Context*: The context is set in an amusement park, which resembles an everyday activity in the real world. The bumper car game is expected to be more familiar for most of students than other events such as a real car collision. *Type of Resources*: No illustrations or any other visual aid.
FM42CI: Contextualized-Illustrated (CI)	
At the amusement park, Jane drives her bumper car into an empty, stopped bumper car. Jane's car now moves slowly, but the empty car moves fast across the smooth, frictionless floor.  Consider the force(s) at the moment when Jane's car hits the empty car. Which statement is correct?	*Type of Abstractness*: Same as the contextualized item. *Nature of Context*: Same as the contextualized item. *Type of Resources*: An illustration is added to guide students' visuals. The illustration not only visualizes the scenario (e.g., Jane and bumper cars), but also shows the direction of the movement, the speed of the motion, and the interaction of forces. For instance, the three horizontal lines next to Jane indicate the speed of Jane's car. The radical lines around where two car contact imply where the forces interact.

Table 2.2
Distractors of item FM42C with targeted misconceptions and calculated proportion of correct response

At the amusement park, Jane drives her bumper car into an empty, stopped bumper car. Jane's car now moves slowly, but the empty car moves fast across the smooth, frictionless floor.

Consider the force(s) at the moment when Jane's car hits the empty car. Which statement is correct?

	Distractors	FM 42C	FCI Q11[1]	Misconception in FCI
A.	Each bumper car exerts a force on the other, but Jane's car exerts a larger force.	0.44	0.50	AR1. Greater mass implies greater force AR2. Most active agent produces greatest force
B.	Jane's car exerts a force on the empty car, but the empty car exerts a larger force back on Jane's car.	0.15	0.07	/
C.	Each bumper car exerts the same amount of force on the other car.	0.12	0.13	Correct Answer [Newton's Third Law]
D.	Jane's car exerts a force on the empty car, but the empty car does not exert a force on Jane's car.	0.29	0.28	AF1. Only active agents exert forces

Note. Distractors for A, C, and CI versions of the same item are almost the same with slight changes in wording. However, the order of distractors is randomized for each of the three versions. The proportion of correct response for each distractor in FM42C is averaged across three context manipulations and two booklets. The proportion of correct response for distractors in FCI is averaged across all student groups reported in FCI. "AR" and "AF" stand for "Action/Reaction Pairs" and "Active Force", respectively, in FCI.

Booklet Design. Given the small sample of students and the number of items we developed (a total of 36 items), it is not feasible to test all items with all variations. We selected items that tapped into the most essential concepts in Forces and Motion and Energy, and make

[1] See Appendix C for Q11 in FCI.

sure each item displays at least two context manipulations in the booklets. Based on these criteria, we created two booklets, each of which had 13 questions on FM and seven questions on E. Each item in the booklets had four possible multiple-choice answers and exhibited one of the three previously defined levels of context richness (A, C, or CI).

Items were distributed into two booklets so that the contextualized (C) version and the contextualized-illustrated (CI) version of the same item did not appear in the same booklet (note that C and CI versions are identical except that the latter is accompanied with an illustration). Items were carefully chosen so that the number of items exhibiting each of the manipulation (A, C, and CI) was approximately uniform across the three contexts. To allow for the comparison of students' performance on two booklets, we incorporated eight common items (four FM items and four E items) in the exact same position of two booklets. The exact ordering of assessment items for each of the two booklets is displayed in Tables 2.3 and 2.4.

Table 2.3
Forces and motion (FM) items in each of the two booklets

	Forces and Motion												
Item #	1	2	3	4	5	6	7	8	9	10	11	12	13
Booklet 1	*31C*	**21C**	11A	**22CI**	*41CI*	**21A**	*32C*	31A	*11CI*	**22A**	*42C*	*12CI*	41A
Booklet 2	*31CI*	**21C**	32A	**22CI**	*41C*	**21A**	*32CI*	12A	*11C*	**22A**	*42CI*	*12C*	42A

Note. Items are coded in the form of *two-digit unique identifier + context level* so that two items with the same identifier cover the same concepts and possibly differ only in the amount of context provided. Note the bolded items are identical in both booklets and the *italicized* items differ only in their context (being either C or CI)

Table 2.4
Energy items in each of the two booklets

	Energy						
Item #	14	15	16	17	18	19	20
Booklet 1	**41C**	**21A**	*31CI*	*11C*	**41A**	11A	**21CI**
Booklet 2	**41C**	**31A**	*31C*	*11CI*	**41A**	31A	**21CI**

Procedure. Students within a period were randomly assigned one of two booklets containing 20 multiple-choice physics questions and given 50 minutes to complete the exam. The randomization was done such that a fixed number of students, approximately half of each period, received booklet 1 ($N = 155$) and the remaining half received booklet 2 ($N = 161$).

2.3.3 Analysis

The data analysis plan of this study is inspired by the work by Kamata (2001) which proves the Rasch model is algebraically equivalent to the hierarchical generalized linear model (HGLM) by treating person ability as a random effect. Building on Kamata's (2001) framework, we explored the effect of contexts at the overall test level and the individual item level. The effects of contexts at the test level is captured by the HGLM model and the item analysis is performed using Rasch models. To ensure the effect of context richness is not confounded with person-characteristic variables, a differential item functioning (DIF) analysis is employed to detect potential gender-related bias in items.

Hierarchical Generalized Rasch Model. Kamata (2001) proposed a two-level HGLM model which is algebraically equivalent to the Rasch model. In his framework the level-1 model is an item-level model and the second level model is a person-level model.

Assume the outcome variable is whether a person j ($j = 1, \ldots, n$) answers an item i ($i = 1, \ldots, k$) correctly, the level-1 structural model can be formulated as:

$$\begin{aligned} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \eta_{ij} \\ &= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \cdots + \beta_{kj}X_{kij} \\ &= \beta_{0j} + \sum_{q=1}^{k} \beta_{qj} X_{qij} \end{aligned} \tag{2.1}$$

and it can be reduced to

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij} = \beta_{0j} + \beta_{qj}X_{qij} \tag{2.2}$$

where p_{ij} is the probability that *j* answers item *i* correctly, $\eta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$ is the logit link function, and X_{qij} represents the *q*th dummy variable (or *q*th item indicator) for person *j*, with values 1 when $q = i$ and 0 otherwise. In order for the design matrix of the model to achieve full rank, one of the dummy variables is dropped and referred as the "reference" item. In this case, β_{0j} can be interpreted as the expected item effect of the reference item for person *j*, while β_{qj} represents difference of an individual item's effect from the reference item, or the specific effect of the *q*th item indicator, namely the effect of the *i*th item for $i = 1,..., k-1$.

The subscript *j* on the βs in the Equation 2.1 indicates that the effects of items are not assumed to be constant across people at level one (item level). This assumption is further addressed in the level-2 model (person-level model) by modeling the intercept β_{0j} as a random effect. In the level-2 model, the intercept β_{0j} is decomposed into two components: the average effect of the reference item and some variations in person ability. The level-2 model is expressed as

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ \quad \dots \\ \beta_{(k-1)j} = \gamma_{(k-1)0} \end{cases} \tag{2.3}$$

where u_{0j} represents the random component of β_{0j}, namely the ability of person *j* ($u_{0j} \sim N(0, \tau)$). Note there is no random terms added to β_{1j} to $\beta_{(k-1)j}$, which suggests that the effects of other items (except for the reference item) are assumed to be fixed across individuals. Combining level-1 and level-2 models, the Equation 2.2 can be reformatted as $\eta_{ij} = \gamma_{00} + u_{0j} +$

γ_{q0} for a specific person *j* and a specific item *i* for $i = q$. Then the probability of a person *j* answers an item *i* correctly can be expressed as

$$p_{ij} = \frac{1}{1 + \exp[-\eta_{ij}]} = \frac{1}{1 + \exp\{-[u_{oj} - (-\gamma_{q0} - \gamma_{00})]\}} \tag{2.4}$$

where $i = q$. Kamata (2001) notes that Equation 2.4 is algebraically equal to the Rasch model,

$$p_{ij} = \frac{1}{1 + \exp\{-(\theta_j - \delta_i)]\}} \tag{2.5}$$

where $\theta_j = u_{0j}$ and $\delta_j = -\gamma_{q0} - \gamma_{00}$. The person parameter θ_j in the Rasch model can be considered either fixed or random variables, while the u_{0j} in the HGLM framework are random variables, with $u_{0j} \sim N(0, \tau)$. In another words, when the person parameter θ_j in the Rasch model is treated as random variables, a HGLM can be considered algebraically equal to the Rasch model.

Note that the two-level HGLM models can also be extended into a latent regression model with person-characteristic variables. The level-1 model remains the same as Equation 2.1, while the intercept at level 2 now takes into account the effect of person characteristics variables:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + \cdots + \gamma_{0p} W_{pj} + u^*_{0j} \tag{2.6}$$

where W_{sj} $(s = 1, \ldots, p)$ are person-level predictor measures for predictor *s* and person *j*. The person ability from the simple Rasch model is now explained by the effects of some person-level predictors and the variations in person ability that are not captured by person-characteristic variables ($u_{0j} = \gamma_{01} W_{1j} + \cdots + \gamma_{0p} W_{pj} + u^*_{0j}$). The two-level HGLM can also be extended into a three-level hierarchical model if students are nested in some larger units (e.g., teachers, classrooms, schools, etc.) or three-level latent regression models by adding third-level predictors.

Test-level Analysis. Building on the work by Kamata (2001), we proposed a HGLM to account for the nested structure among items and students using the generalized linear mixed modeling (GLMM) framework. To detect whether there is between-student variability, we compared two unconditional models with sequential entry of random effects. The first model (*m1*) models the outcome variable as a function of the intercept with no random effects. The second model (*m2*) adds a second-level random effect of students. Comparing *m1* with *m2*, the likelihood ratio test yields no significant difference between the two models ($\chi^2(1) = 0.27$, $p > 0.05$), suggesting that the variability among students may not be different from what we would expect to see due to random chance. However, from the perspective of study design, we decide to keep the two-level structure because student is an important factor when considering item clustering and interpreting model results. Besides, the two-level model enables us to examine the effect of context richness on students' performance taking into account the potential impacts of item covariates and eliminating the variability due to individuals as much as possible. All models were estimated using the *lme4* package in *R*.

Item-level Analysis. As the HGLM offers overall interpretation on the effect of context richness on students' performance, we employed Item Response Theory (IRT) models to gain deeper insights into how three levels of context richness are associated with the item difficulty of individual items. The purposes of the item-level analysis are to check (1) whether the effect of context richness detected at the test level retains for most of individual items, and (2) whether there is additional information regarding the context effect that is not captured at the test level.

Three IRT models were fitted to the data: Rasch model, one-parameter logistic (1PL) model, and two-parameter logistic (2PL) model. Model comparison were based on three absolute fit indices: AIC, BIC, and -2log-liklihood. As three models performed very closely in all three

indices, the most parsimonies model—Rasch model—was selected given our sample size and its interpretability. To fulfill the unidimensionality assumption of the Rasch model, models for FM items and E items were estimated separately. To examine whether 13 FM items loaded onto one factor and whether 7 E items loaded on the other factor, a two-factor confirmative factor analysis (CFA) was conducted for booklet 1 and booklet 2, respectively. The results of CFAs reported a root mean square error of approximation (RMSEA) smaller than 0.08 (0.076 for booklet1 and 0.052 for booklet 2), indicating a good to acceptable fit (MacCallum, Browne, & Sugawara, 1996). The item fit statistics for the Rasch models are presented in Appendix A. As we administered two different booklets to different students, two groups' performance were equated using the simultaneous calibration approach (de Ayala, 2009). In this approach, two samples' response data are concatenated and calibrated in one analysis through the link of common items. The *ltm* package in *R* was used to estimate all IRT models.

Checking for Gender Bias. To make sure the results obtained from the previous analysis are not confounded with students-characteristics variables, we performed a differential item functioning (DIF) analysis[2] with gender. Gender is our primary variable of interest as prior studies have documented male students tend to outperform female students in science assessments (Bayraktar, 2009). The other two variables (ELL and SPED) are not examined given their small sample sizes and the fact that both predictors showed a sign of perfect separation.

We performed DIF by fitting a logistic regression to each item. Benjamini-Hochberg adjusted *p* value correction is used to determine significance with multiple comparisons (Mcmartin, Mckenna, & Youssefi, 2000). Results (see Appendix B) confirm that no items

[2] DIF analysis was performed with only 239 students (out of 316 students) due to an absence of 77 students' demographic information, which means the results of the gender bias analysis may not be generalized to those whose gender are unknown in our study. Despite the small sample size, we believe this is an essential step to validate our results and make inference about context effects.

function differently for students with the same latent ability but different gender. These results provide evidence to the claim that the variation observed in students' performance among different items is not due to the gender difference.

2.4 RESULTS

This section starts with a general description of students' performance in the test. To examine the effects of adding context and adding illustrations to assessment items, we first present the results of HGLM, which offers a test-level interpretation of the effects of context richness. Next, we discuss the findings from the Rasch models, which provide item-level evidence to support and supplement the results reported by HGLM.

2.4.1 General Description of Student Performance

The mean score on the instrument across two booklets is 0.311 (sd = 0.463). The average score of booklet 1 (M = 0.314, sd = 0.464) is almost equivalent to the average score of booklet 2 (M = 0.308, sd = 0.461). The two-sample t-test suggests that students who received booklet 1 did not perform significantly different from students who took booklet 2 on common items, $t(2523) = -0.058$, $p > 0.05$. Although the test results seem not to be ideal, it is consistent with scores reported by other similar studies which intentionally tap into students' misconceptions. For example, Bayraktar (2009) reported a mean score of 40.89% (sd = 0.12) on FCI in his study. Hestenes et al. (1992) reported similar mean scores for high-school student groups, ranging from 27% to 73% (the average mean score across groups is 37.38%). Clement (1982) also found 88% of students answered questions incorrectly when asked about forces at the vertical direction. We calculated the proportion of correct response for the most attractive distractor in each item (see

Table 2.5). As expected, distractors that are most appealing to students are those strong misconceptions widely documented in the literature.

Table 2.5
Proportion of correct response for the most attractive distractor in each item

Items	Overall Proportion of Correct Responses	Proportion of Correct Response for the Most Attractive Distractor	Code in FCI	Misconception	Source of Assessment Items
FM11	0.48	0.20	Ob	Obstacles exert no force	FCI
FM12	0.15	0.71	R2	Motion when force overcomes resistance	FCI
FM21	0.28	0.37	AF7	Active force wears out	FCI
FM22	0.15	0.49	AF7	Active force wears out	FCI
FM31	0.17	0.38	AF2	Motion implies active force	FCI
			AF7	Active force wears out	FCI
FM32	0.21	0.54	AF2	Motion implies active force	FCI
FM41	0.22	0.47	AR1	Greater mass implies greater force	FCI
			AR2	Most active agent produces greatest force	FCI
FM42	0.12	0.44	AR1	Greater mass implies greater force	FCI
			AR2	Most active agent produces greatest force	FCI
E11	0.48	0.21	/	Energy not associated with motion	Kruger (1990)
E21	0.34	0.28	/	Energy is not conserved	Kruger (1990)
E31	0.37	0.37	/	Energy confused with force	Kruger (1990)
E41	0.70	0.15	/	Energy confused with force	Kruger (1990)

Note. The overall percentage of correct response and the proportion of correct response for the most attractive distractors are averaged across three context manipulations and two booklets in the instrument.

2.4.2 HGLM Model

A two-level model is specified to account for the hierarchical structure of items being nested in students. We hypothesize that students' performance on a particular item may be affected by which booklet they receive, topic of the item (FM or E), and richness of the context (A, C and CI). Given the students' varying performance on the topics (FM and E), it is also hypothesized that there may be an interaction between topics and the richness of the context. Specifically, the final model is expressed as

$$p_{ij} = \gamma_{00} + \beta_1(Booklet)_{ij} + \beta_2(Topic)_{ij} + \beta_3(Context)_{ij} + \beta_4(Topic * Context)_{ij} + u_{0j} + r_{ij} \quad (2.7)$$

where p_{ij} refers to the probability of person *j* responding to item *i* correctly, γ_{00} is the level-2 intercept, $\beta_q(q = 1, \dots, 4)$ represents the effect of the *q*th predictor, u_{oj} denotes the ability of person *j*, r_{ij} is the random error assicated with the item *i* nested within person *j*.

Table 2.6 presents the output of the final model. The overall model fit is significantly better than the null, $\chi^2(6) = 491.29$, $p < 0.001$. Only a small portion of variances in the outcome are explained by the level 2 random effect (*variance* = 0.03), which is consistent with our analysis with unconditional models. As indicated in Table 2.6, topic was uniquely predictive of students' performance (b = -1.16, *se* = 0.09, $p < 0.001$) after controlling for booklet, context and random effects. Taking Forces and Motion (FM) items decreases the probability of getting an item correct by 1.16 logits compared to taking Energy(E) items, corresponding to a probability of 0.24. Meanwhile, significant effects of context richness are observed at $\alpha = 0.05$. Compared to abstract items, contextualized items are associated with an improvement in test performance by 0.38 logits (*se* = 0.10, $p < 0.001$), corresponding to a probably of 0.59; while contextualized-

illustrated items are related to a decrease in test performance by 0.47 logits ($se = 0.11$, $p < 0.001$), associated with a predicted probability of 0.38.

Finally, there was also a significant interaction between topic and context on test performance. To understand the nature of the interaction, predicted values were plotted for each topic by three context manipulations (A, C, CI). As illustrated in Figure 2.1, the positive effect of contextualized items is more pronounced under the Energy (E) topic than under the Forces and Motion (FM) topic ($b = -0.34$, $se = 0.14$, $p < 0.05$). In other words, adding contexts to an item leads to a greater probability of correct answer when contextualized items are used with the Energy topic rather than with the Forces and Motion topic. On the other hand, adding illustration to contextualized items makes items more difficult under Energy topic than under the Forces and Motion topic ($b = 0.30$, $se = 0.14$, $p < 0.05$).

Table 2.6
Coefficient estimates of HGLM

	B	***SE***
Fixed Effect		
Intercept	-0.05	0.07
Booklet (BK2)	-0.03	0.06
Topic (FM)	-1.16 ***	0.09
Context (Contextualized)	0.38 ***	0.10
Context (Contextualized + Illustration)	-0.47 ***	0.11
Topic (FM) x Context (C)	-0.34 *	0.14
Topic (FM) x Context (CI)	0.30 *	0.14
Random Effect	***Variance***	***SE***
Student	0.03	0.19

Note. All coefficient estimates are in logits. The outcome variable is dummy coded with 1 as correct and 0 as incorrect. Booklet is dummy coded using booklet1 as the reference. Topic is dummy coded with Energy being the reference group. Contexts are dummy coded with Abstract (A) items as the reference group. Number of students = 316.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

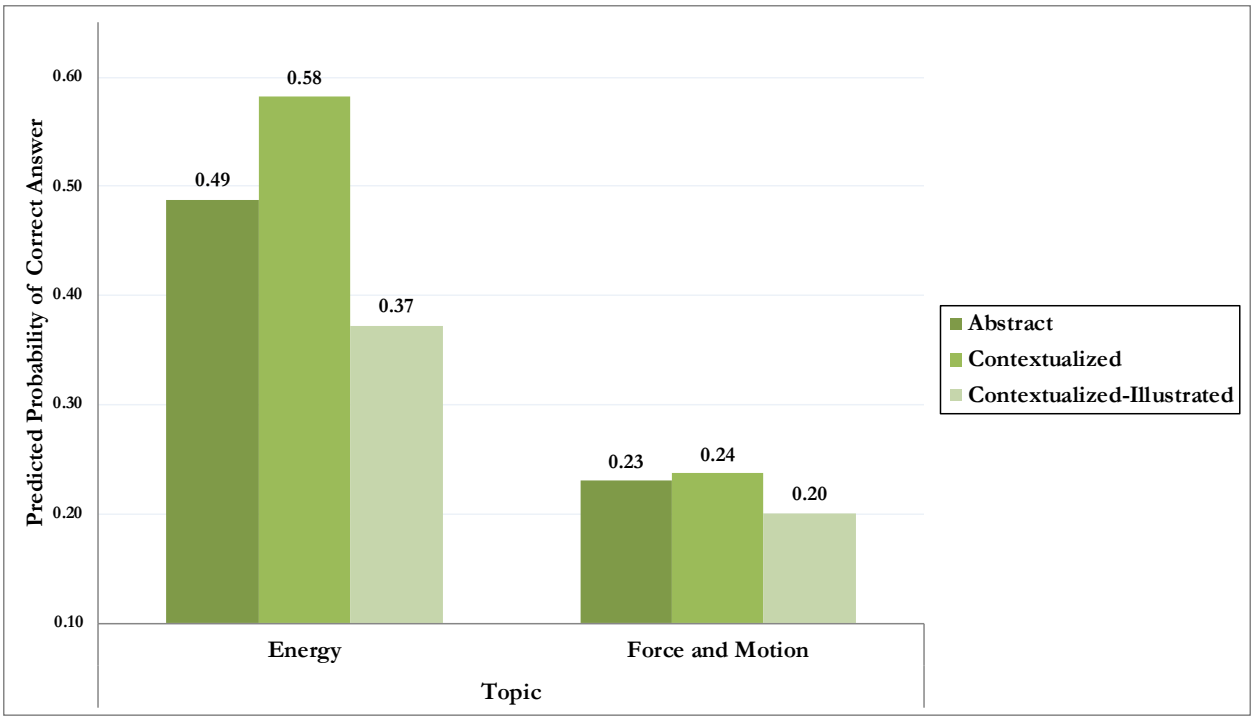


Figure 2.1 Interaction between topic and context.

2.4.3 Rasch Model

As the hierarchal model uncovers the overall differences in students' performance due to topics and the richness of contexts, the results of Rasch models indicate that the effect of context richness varies on the item base. Table 2.7 presents the item parameters estimated by the Rasch models for FM and E items. On average FM items seem to report higher item difficulty than E items despite context manipulations. Figure 2.2 and 2.3 present the item characteristic curves (ICC) and test information curves for all items. As indicated in Figure 2.3, E items have a better performance rating when discriminating students with average ability, while FM items tend to be more informative when diagnosing students with a higher latent ability range. Among all items, FM11 and E11 report item difficulty that are not significantly different than zero. Those two items are intentionally designed to be easier as confidence boosters to prevent students from giving up in the middle of the test because the test is too difficult.

Table 2.7
Item parameters for the Rasch model for FM and E items

Item ID	Context	Item Difficulty	SE	Z Values
FM11	A	0.22	0.19	1.16
	C	0.21	0.18	1.12
	CI	-0.11	0.19	-0.58
FM12	A	1.72	0.23	7.60
	C	2.25	0.26	8.57
	CI	2.07	0.25	8.12
FM21	A	0.93	0.14	6.61
	C	1.21	0.15	8.23
FM22	A	1.64	0.16	10.20
	CI	2.25	0.19	11.87
FM31	A	1.47	0.22	6.73
	C	1.70	0.23	7.37
	CI	2.19	0.26	8.49
FM32	A	1.67	0.22	7.48
	C	1.65	0.23	7.25
	CI	1.20	0.20	5.91
FM41	A	2.07	0.25	8.12
	C	1.20	0.20	5.91
	CI	1.23	0.21	5.91
FM42	A	2.13	0.25	8.41
	C	1.90	0.24	7.82
	CI	2.79	0.32	8.82
E11	A	0.15	0.19	0.79
	C	0.06	0.19	0.30
	CI	0.10	0.19	0.52
E21	A	0.73	0.14	5.23
	CI	0.89	0.14	6.20
E31	A	0.68	0.19	3.52
	C	0.46	0.19	2.42
	CI	0.76	0.20	3.85
E41	A	-0.92	0.14	-6.40
	C	-1.04	0.15	-7.14

Note. Item discrimination is set to 1 for all items in the Rasch model.

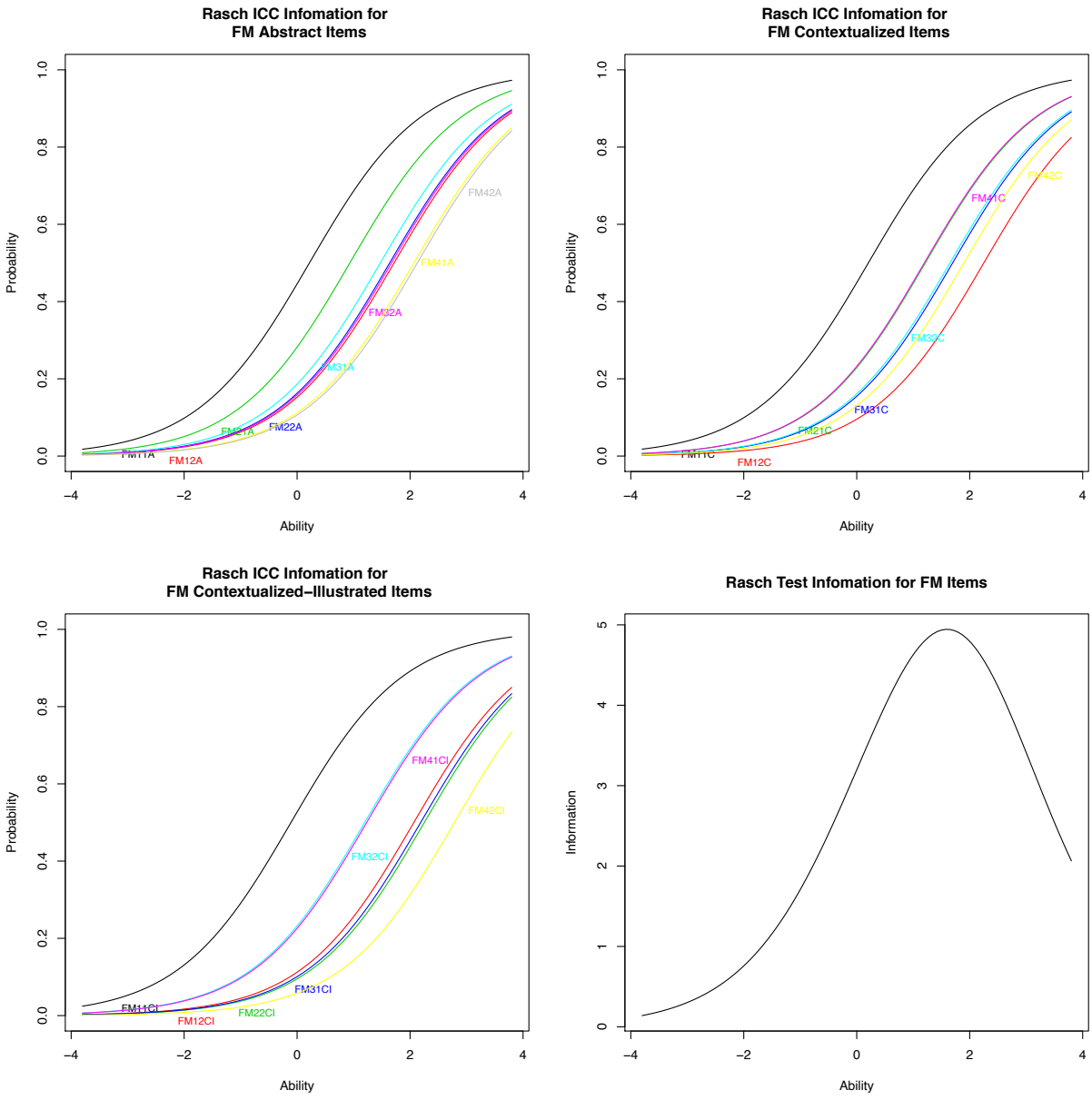


Figure 2.2 Item characteristic curves and test information curve for FM items.

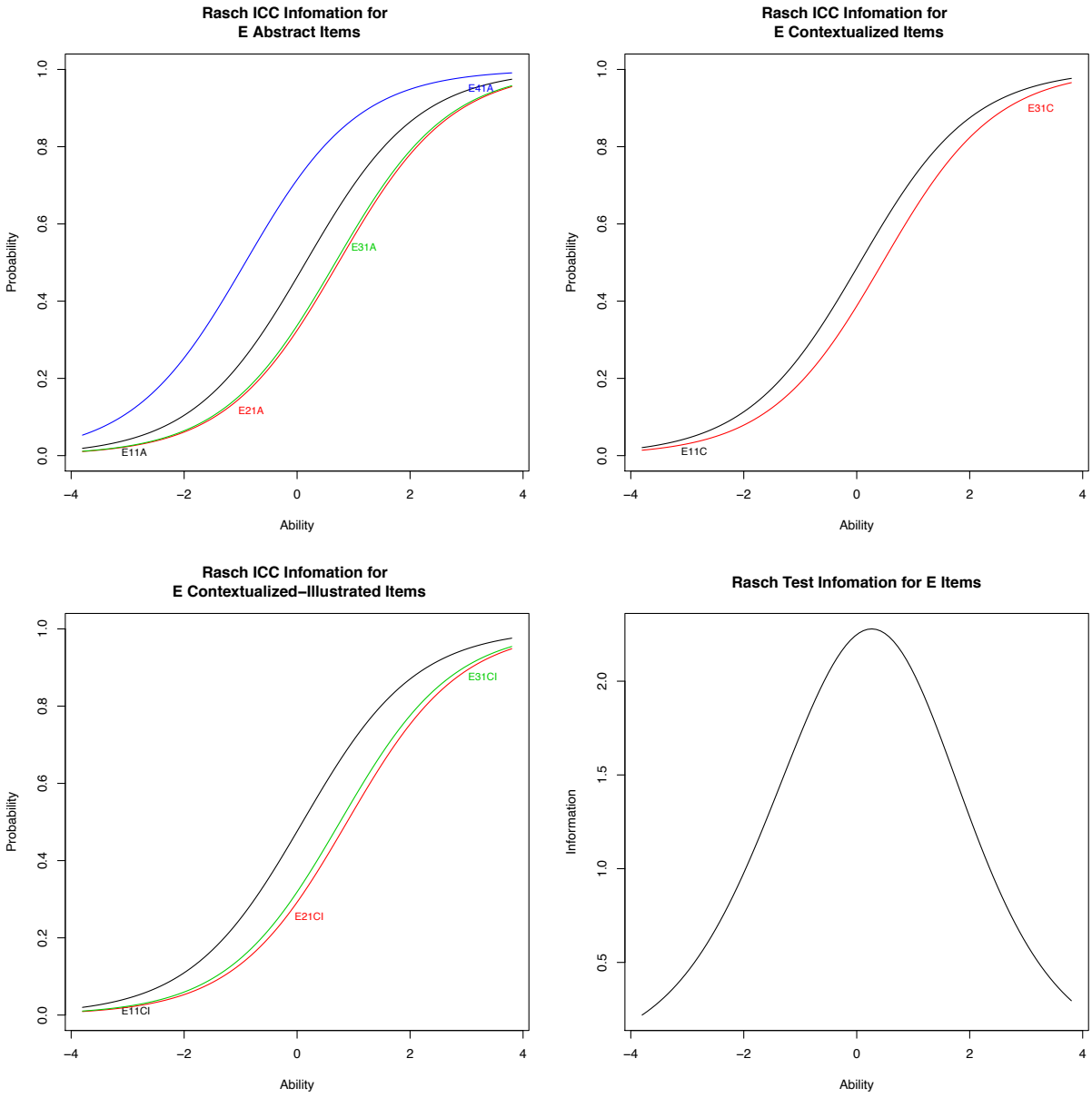


Figure 2.3 Item characteristic curves and test information curve for E items.

To understand the effect of the three levels of context richness, we compared the differences in item difficulty among each two of three context manipulations for each item. According to Table 2.8, 70% of items (four out of seven FM items and three out of three E items) showed that adding contexts made the item easier compared to its abstract version (A versus C). Five out of eight items in total as 63% suggested that illustrations added difficulty to contextualized items (C versus CI). Comparing abstract items against contextualized-illustrated items (A versus CI), 60% of items (four out of seven FM items and two out of three E items) demonstrated an increase in item difficulty when items were contextualized and illustrated. Scrutinizing through Table 2.8, we find the magnitude of difference in item difficulty was greatest between the A-CI comparison (average difference across all items = -0.068), followed by C-CI (average difference = -0.062) and A-C (average difference = 0.041) comparisons, which is consistent with the results of HGLM (Table 2.6 and Figure 2.1).

Beyond the overall discussion of item performance, we also observed a potential association between the cognitive demand of items due to content difficulty and the magnitude of context effect. The cognitive demand of an item is roughly characterized as the item difficulty reported by the Rasch models averaged across three variations[3]. We refer it as the *average item difficulty* for brevity in the subsequent discussion. The average item difficulty is used as a proxy of content difficulty and is presented in Table 2.8. We acknowledge that the characterization of cognitive demands is over-simplified, and the low item difficulty could be due to the effect of context manipulation rather than cognitive loads posed by content. However, we argue that if an item maintains a relatively low item difficulty regardless of context manipulations, it is likely

[3] Item difficulty is not available for some item variants due to booklet design. For those items, the average item difficulty is averaged across whatever variants of the item that are available.

that the source of item difficulty comes from the cognitive demands imposed by the physics concepts assessed in addition to item contexts.

Table 2.8
Differences in item difficulty for FM and E items

Item ID	Item Difficulty			Average Item Difficulty	Difference in Item Difficulty		
	A	C	CI		A - C	C - CI	A - CI
FM11	0.22	0.20	-0.11	0.10	0.02	0.31	0.33
FM12	1.72	2.25	2.07	2.01	-0.53	0.18	-0.35
FM21	0.93	1.21	-	1.07	-0.28	-	-
FM22	1.64	-	2.25	1.95	-	-	-0.61
FM31	1.47	1.70	2.19	1.79	-0.23	-0.49	-0.72
FM32	1.67	1.65	1.20	1.51	0.02	0.45	0.47
FM41	2.07	1.20	1.23	1.50	0.87	-0.03	0.84
FM42	2.13	1.90	2.79	2.27	0.23	-0.89	-0.66
E11	0.15	0.06	0.10	0.10	0.09	-0.04	0.05
E21	0.73	-	0.89	0.81	-	-	-0.16
E31	0.68	0.46	0.76	0.63	0.22	-0.30	-0.08
E41	-0.92	-1.04	-	-0.98	0.12	-	-

Note. Item difficulty is not available for some item variants due to booklet design.

To investigate the potential association between content difficulty and the magnitude of context effect, a Pearson correlation was conducted between the average item difficulty and the difference in item difficulty between every two variants of an item. Table 2.9 presents the result. Based on the correlation coefficients, content difficulty is positively associated with the difference in item difficulty across all three groups of comparison (A vs. C, C vs. CI, and A vs. CI), especially with the difference between abstract items and contextualized-illustrated items (A vs. CI). In other words, the increase in content difficulty is associated with the larger magnitude of context effect, regardless of whether it is a positive or negative effect.

Table 2.9
Correlation between average item difficulty and the difference in item difficulty between every two variants of an item

	A-C	C-CI	A-CI
Average Item Difficulty	0.44	0.52	0.73

To support the claim, we take a closer look at all the items in Table 2.8. For items that are less demanding in terms of content difficulty, such as F11 (average item difficulty = 0.10), F32 (average item difficulty = 1.51) and E11 (average item difficulty = 0.10), the difference between the item difficulty of abstract versions and the item difficulty of contextualized versions is almost neglectable. This shows the positive effect of adding contexts is less salient or even neglectable for items with low content difficulty. In contrast, when items are relatively demanding, the negative effect of overcomplicated context (contextualized and illustrated) become more notable. This is proved by FM22 (average item difficulty = 1.95), FM31 (average item difficulty = 1.79), FM42 (average item difficulty = 2.27), E21 (average item difficulty = 0.81) and E31 (average item difficulty = 0.63), which all report contextualized-illustrated versions as the most difficult version. Although the present study reveals the potential association between content difficulty of an item and the effect of the two item characteristics, the observation is unsystematic. Further work is needed to provide a systematic and rigorous investigation into the relationship.

2.5 DISCUSSION

This study explored the effect of context richness, including textual and visual item characteristics, on students' performance in a middle-school physics assessment. Results of the test-level analysis suggest that using contextualized items improves students' test performance

compared to using abstract items, while adding illustrations in addition to contexts does not seem to bring in additional benefits. In fact, contextualized-illustrated items lead to poorer performance compared to students' scores using contextualized items, or even abstract items. Furthermore, a joint effect of contexts and topics was detected. Using contextualized items led to greater increase on performance under the Energy topic than under the Forces and Motion topic. Similarly, adding illustration to contextualized items makes items more difficult under Energy topic than under the Forces and Motion topic. Results suggest that perhaps certain topics are more compatible with context manipulation, while for other topics the effect of context richness may not be as obvious.

While the test-level analysis uncovers a clear pattern of context effects on students' overall performance, the item-level analysis further suggests that the effect of contexts vary by items. The variations can be partially attributed to multiple factors such as the complexity of contexts (combinations of item characteristics), content difficulty (cognitive loads due to the intended science concepts), and learner differences. In the present study we observed a potential relationship between the content difficulty and the magnitude of context effects, which calls for further investigation.

As the textual advantage of contextualized items has been well documented in the literature, we believe the poor performance associated with contextualized-illustrated items detected deserves more discussion. Given that CI items have the richest contexts among three kinds of items, it is reasonable to infer that an overcomplicated item context may not necessarily be beneficial for students. In this section, we further discuss how adding illustrations may make the context unnecessary complex, thereby negatively impacting students' performance. One factor that may account for the low performance is the fact that students' attention is split

between the two modes of representation (Berends & van Lieshout 2009). It is especially true when visual representations do not coordinate with the textual information, more cognitive resources need to be allocated to process split-attention materials, thus reducing the resources available for learning and problem solving (Cook, 2006). Such "split-attention effects" may slow down the comprehension process (Brookshire, Scharff & Moses, 2002) and result in heavy extraneous cognitive load, especially for novice learners (Cook, 2006).

Another possible explanation to students' poor performance may be that adding illustrations to the context may create additional "noises", such as irrelevant details and misleading information, which drives students away from what is intended from the question and lead to the activation of inaccurate schema. According to Ahmed & Pollitt (2000), any irrelevant information presented in illustrations may cause students to focus on the context rather than the content of the item. This may cause an over-interpretation or misinterpretation of information presented in illustrations. Furthermore, it may lead students to fall into their everyday knowledge schema rather than the targeted scientific principles to solve the problem. To further unfold this idea, we argue that adding illustrations may make the item so "realistic" that it triggers students' alternative conceptions evolved in their daily life. As Bayraktar (2009) points out, misconceptions are alternative conceptions that "mainly originate from person's experiences with the real world that seem very logical to them" (p. 273). Strong visual cues may help provide a better simulation of realities and reproduce a scenario that is close to what students encounter in everyday life. When students fall onto their everyday knowledge schema to solve problems, it may trigger misconceptions (or alternative conceptions) rooted in their personal experiences, lowering their performance in the test.

Finally, we present a hypothesis that illustrations may add another layer of cognitive loads to students by assessing practical problem-solving skills in additional to science concepts. One strong argument favoring contextualized items is that it allows the assessment of transfer of learning by linking school instruction with real-world problems (Boaler, 1993; Wang, 2016). In this case, the illustrated context measures not only the depth of students' scientific understanding, but also their ability to decode illustrations, filter relevant information, make referential link to real-world situations, and apply knowledge to explain various contexts, etc. On the contrary, in abstract items students may easily find a pattern or clue of what the question is asking and simply apply test-taking strategies to solve the problem without the need for grappling with the context.

With the current study, we cannot rule out either of the interpretations (and there are more possibilities that we did not capture). Therefore, we propose that cognitive interviews with students may be necessary to trace the reason of students' poor performance on the contextualized-illustrated items and unfold the complexity of contexts. However, it seems safe to conclude that all factors point to the importance of creating high-quality of illustrations that are relevant to the target science concepts and coordinate with the textual mode of information.

Beyond the interpretation of context effect at the test level, results of the item-level analysis reveal an overall trend that is consistent with the textual advantage of contextualized items and the challenges of adding illustrations in addition to contexts detected at the test level. Results further suggest that the magnitude of context effect on students' performance varies depending on the cognitive load provoked by the content of the item. Specifically, an increase in the content difficulty of an item amplifies the difference in item difficulty between any two variants of the item. The magnitude of difference is most substantial between abstract (A) and

contextualized-illustrated (CI) items, followed by the comparison between C and CI and then A and C. In other words, while contextualized items seem to be generally beneficial for students, including context to an item assessing complex science concepts (and thereby are more cognitive demanding in terms of content) leads to greater improvement on students' performance compared to adding context to an item with low content difficulty. Similarly, when an item is content demanding, adding illustrations in addition to context leads to greater decrease on students' performance in contrast to students' performance on a CI item measuring easy science concepts.

This pattern raises an important discussion about the balance between the cognitive load caused by item characteristics and the cognitive demands due to the underlying science concepts being measured. Theorists of cognitive load theory claims that students have limited work memory (Cook, 2006). Item characteristics should supplement the understanding of intended science concepts, not compete with it. Our previous discussion has demonstrated how sophisticated the thinking process could be when navigating through a contextualized or even contextualized-illustrated item. When the item itself requires a high-level understanding of underlying science concepts, an over-contextualized item (e.g., CI items) may cause students to get overwhelmed or distracted by the details presented in the contexts and overlook the intended science knowledge structure. This aligns with findings by Mevarech and Stern (1997) that students perform better on math problems embedded in sparse contexts than real contexts. Mevarech and Stern (1997) argue that context characteristics tended to drive students' attention from the mathematical features and activate a simplistic mathematical model rather than a holistic mental representation of the abstract mathematical concepts. On the contrary, rich contexts (e.g., illustrations) could become an asset when the item requires low-level scientific

thinking. As more cognitive resources are available for learning and information processing, contextualized items, if well-presented, could help activate prior knowledge and comprehend the text better and faster.

Besides the association between content difficulty and context effects, we also explored some common item characteristics embedded in each item that could potentially lead to the discrepancy in students' performance across three context manipulations. We scrutinized through item characteristics such as familiarity of the setting (e.g., driving a bumper car versus playing basketball in a wheelchair), agents involved in the contexts (e.g., active versus passive agents) , motion indicated in the item (e.g., slow down versus speed up) and the direction of forces featured in the context (e.g., horizontal forces versus vertical forces). No consistent patterns were observed in terms of students' performance which could potentially reveal the associations between three context manipulations and any of the subtle item characteristics described above. We speculate that this may be because each item represents a unique combination of item characteristics and taps into different science concepts and misconceptions. However, the limited number of items makes the present study under-powder to detect factors that influence the effect of context richness at such a subtle level.

Our observation is echoed by Ahmed and Pollitt (2000) that students react in "an individual and unpredictable way" (p. 1) to various contexts. This brings in implications that students' individual differences need to be taken into account when interpreting the effects of contexts on student performance. For instance, certain contexts may be more familiar for some students. Some students may have better visual-spatial ability to process illustrations betters, while others may be more comfortable with text-only format; some contexts may impose different interpretations due to cultural differences.

So far we have presented two approaches (test-level and item-level) to understand the effect of context richness on students' test performance. Results of two approaches present different aspects of the study but both highlight the fact that whether or not to use contexts is not a simple yes-or-no question. It may seem that using contextualized items introduce additional variances which we do not know its influence and are difficult to control. However, we argue that by detecting and ruling out part of the variances we can take control of the inquiry process. In fact, teachers may get empowered by knowing more about the underlying mechanism of context effects, as it enables them to think about the trade-off between context and content complexity when designing items, choose items based on a specific purpose (e.g., to measure students' ability to transfer certain knowledge across certain contexts), and make meaningful interpretation on the performance of a specific student group.

This work can be extended in two directions. First, constrained by the sample size, the present study presents the two variants of an item in one booklet (either A and C or A and CI). As variants of an item are designed to assess the same physics concepts but differ in the level of context richness, exposing students to variants of the same item in one booklet may cause a problem similar to the testlet effect. More specifically, students' responses to the first variant may impact their responses to the next variant as they may recognize the two items are similar. In the current study, we checked for the testlet effect using Q_3 statistic proposed by Yen (1984), which computes the correlation matrix of residuals and looks at the maximum value. Based on the critical value of 0.20 (Chen & Thissen, 1997), the correlations of residuals for all item sets (a set contains three variants of an item) did not exceed the critical value, indicating minimal local dependence, if any, among variants of an item. In the future work, we will improve the experiment design to measure the effect of the two item characteristics more accurately. For

example, we will try to include each of the three variants only in one booklet, which may eliminate the potential influence of local dependence among variants of an item. Second, as we detect a potential association between content difficulty and the magnitude of context effect, we need to take a closer look at the association. With more items and careful manipulation of context characteristics, we will investigate how context effect is mediated by the difficulty associated with the underlying constructs, such as content difficulty, and further explore statistical procedures to account for the content difficulty if any (e.g., an interaction between context manipulation and content difficulty).

Chapter 3

Exploring the Role of Context Familiarity on Science Assessment Outcomes: A Cognitive Diagnostic Modeling Approach

Dongsheng Dong[1], Min Li[1], Philip Hernandez[2], Jim Minstrell[3]

Klint Kanopka[2], & Maria Araceli Ruiz-Primo[2]

[1] College of Education, University of Washington

[2] Graduate School of Education, Standard University

[3] Facet Innovations

Abstract

Context familiarity refers to students' prior experiences with the object(s) of interest described in the scenario of an individual item or in the general context shared by a cluster of item bundle. This study explores the effect of context familiarity, indexed by daily life experience and classroom experience, on two assessment outcomes using a cognitive diagnostic modeling (CDM) approach. Instead of providing a general interpretation of whether the context familiarity impacts test performance, this paper aims at offering a deeper insight into *how* it is related to assessment results in terms of students' attribute profiles and CDM item parameters. The G-DINA (generalized deterministic inputs, noisy "and" gate) model and reduced models were applied to analyze 1,478 students' responses to a contextualized physics assessment and eight post-test survey questions regarding students' familiarity with the context presented in assessment items. Results of the study suggest a potential association between context familiarity

and the probability of mastering certain attributes, which requires further investigation. Specifically, if students are exposed to a similar context before in the classroom (and thereby are familiar with the context of the assessment item), their probability of mastering the physics concepts assessed by the item may be impacted, either positively or negatively. Similarly, having classroom experience (and thus are familiar) with the item context reported a small effect size on reducing the probability of guessing, which refers to the probability of answering an item correctly without understanding any required physics concepts.

3.1 INTRODUCTION

The use of context in assessment items have been highly recommended by research and current reform documents (e.g., Almuna Salgado, 2017; National Science Education Standards, 1996). According to the National Science Education Standards (1996), assessment tasks should be presented in an authentic context that is similar to tasks in which students engage in their daily lives. The incorporation of context provides an inspiring way to connect school learning with real-world practices. However, as a context ensembles multiple context characteristics, researchers express the concern that the influence of some item characteristics on students' performance may not be clearly understood (De Lange, 2007). Context familiarity is one of those item characteristics.

An assessment item is familiar to students when it is connected to students' prior knowledge, which includes the information, knowledge, emotion, experience, and cultural awareness stored in students' knowledge system and activated by the present text (Lee, 2011). Research on familiarity has mostly concentrated on content (Lee, 2011), such as topic familiarity, and context (Ahmed & Pollitt, 2007; Crisp, 2011) such as settings, phenomena and culture presented in the context (Wang, 2016). While there is relatively large body of research on familiarity with content (e.g., topic, Freebody & Anderson, 1983), empirical studies on context familiarity remains sparse (Almuna Salgado, 2017). The current study presents an attempt to provide empirical evidence on the influence of context familiarity on students' performance and item statistics.

We define *context familiarity* as students' prior experiences with the *object(s) of interest* described in the scenario of a stand-alone science assessment item or in the general context shared by a cluster or a bundle of items. The object(s) of interest refers to the object(s) that the

item/cluster focuses on. For physics contextualized items, especially items targeting the topic of Forces and Interactions[4], the object(s) of interest is usually presented in an authentic context that allows students to infer about its motion or related forces explicitly or implicitly. For instance, students may be asked to explain the forces applied on a school bus when the school bus is at rest to pick up students. In this case, the school bus (as well as its driver and students) is considered the object of interest, situated in a general school-related setting of the bus picking up students. To operationalize the concept of context familiarity, the present study focuses on two types of individual experiences as the source of familiarity: *daily life experience* (whether students have seen, used, and/or interacted with the object(s) of interest depicted in the context in their daily lives), and *classroom experience* (whether students have used, discussed and/or interacted with the object(s) of interest presented in the context in classroom practices, such as experiments, group discussions, etc.).

The purpose of the present study is twofold. The primary focus of this paper is to investigate the effect of context familiarity, indexed by daily life and classroom experience, on science assessment results. The two types of experiences were characterized through a set of survey questions at the end of the test. Students' responses to the survey questions were then used as independent variables to predict two important aspects related to assessment results: students' performance and item statistics estimated from the statistical models. Specifically, we focus on individual students' mastery pattern of assessed physics concepts (referred as "attribute profiles") as an indicator of students' performance. Our interest in item statistics lies on the two item parameters estimated from the cognitive diagnostic models (CDM), guessing and slip, which will be elaborated in the Method section. We frame our research questions as follows:

[4] "Forces and Interactions" is referred as "Forces and Motion" in Chapter 2.

1. Does context familiarity have an effect on students' mastery of cognitive attributes assessed in the test items? If yes, which source of familiarity is associated with the effect?
2. Does context familiarity have an effect on the probability of guessing and slip of test items? If yes, which source of familiarity is associated with the effect?

Besides the interest in the item characteristic, the second goal of this paper is to demonstrate the application of cognitive diagnostic models (CDM), specifically the G-DINA (generalized deterministic inputs, noisy "and" gate) model and reduced models, to provide diagnostic feedback about students' learning and statistics of assessment items. The application of CDM encompasses two major methodological focuses. One is to interpret students' mastery of the eight physics-related attributes based on the CDM results. The other focus is to explore how CDM parameters can be linked to students' survey responses about context familiarity. As few studies, if any, have touched upon this issue, this study represents an attempt to provide an exploratory approach to utilizing CDM results for test interpretation and advance the psychometric methodologies used in the research of contextualized items.

3.2 BACKGROUND

This section begins with an overview that discusses the current research on context familiarity in science assessment items regarding its advantages and challenges in empirical applications. The overview is followed by an introduction of the general cognitive diagnostic modeling framework, a statistical approach we choose to employ to address the issues revealed in the discussion of context familiarity above. Strong attention is given to a set of specific cognitive diagnostic models—the generalized deterministic inputs, noisy "and" gate model (G-DINA; de la Torre, 2011) and special cases of the G-DINA model. We discuss the psychometric

properties of the models and describe how they are applied in this study to address research questions.

3.2.1 Context Familiarity

Extensive research has been conducted on the role of context familiarity on STEM assessments. Across various studies, different aspects or sources of context familiarity were explored, such as instructional experiences (Li, Ruiz-Primo, Wills, & Giamellaro, 2012b), cultural differences (Song & Bruning, 2016), education background (Anderson, Reynolds, Schallert, & Goetz, 1977), daily activities (Boaler, 1994), and familiarity of vocabulary (Crisp, 2011). In the subsequent sub-sections, we first present three studies in detail on how they conceptualize context familiarity. Then we discuss the advantages and concerns of setting assessment items in familiar contexts based on empirical evidence collected from prior work.

The Role of Context Familiarity on Students' Performance. Anderson et al. (1977) conducted an experiment on how students' different education background affected comprehension. They hypothesized that individuals' different schema may turn them to interpret texts in certain ways. Participants were 60 college students with background in physical education (n = 30), specifically wrestling, and music education (n = 30), respectively. In the study, participants read two passages, one was related to wrestling and the other was related to music. Each of the passages was constructed in a way that enables two distinct interpretations of the content. After reading the passages, students took 10 multiple-choice questions that had two correct answers corresponding to the two interpretations. Results showed students reported better performance on passages and questions that were related to their education background.

In another study, Li, Ruiz-Primo, Wills, and Giamellaro (2012b) operationalized context familiarity as how closely an item context reflected and related to the instructional activities that

students had experienced when they were learning a given science module. Guided by their definitions, two types of assessment items were developed based on the *proximity* of assessment items to the enacted curriculum: close and proximal. A *close* item aligned with the content described in the curriculum and presents a setting that was similar to the daily activities implemented in the classroom. A *proximal* item measured knowledge and skills relevant to the curriculum, but the context differed from what was studied in the given science module. The proximal items were further divided into proximal 1 (near proximal) and proximal 2 (far proximal) given how much change was applied to the item based on the module. P1 indicated small changes and P2 reflected more substantial changes. Figure 3.1 shows an example of close, proximal 1 and proximal 2 items in a question set[5]. In the study, researchers examined the relationship between item proximity and students' gain from pretest to posttest. It was found that close items yielded larger learning gains from pretest to posttest compared to proximal items, and the magnitude of learning gain decreased as items became more distal from the curriculum.

A recent study by Song and Bruning (2016) focused on another source of context familiarity, cultural background, and investigated how it impacted comprehension, recalling and cognitive load. Technical information regarding global warming was presented in two geographical contexts: a U.S. geographical setting and a South Korean setting. As participants were 147 US college students, researchers posited that students were more familiar with the US setting than the South Korean context. Figure 3.2 shows an example of the two different settings. The findings of this study confirmed their' hypothesis. Students demonstrated deeper learning with respect to application of information in making inferences about global warming and

[5] The example was selected from another work by Li and her colleagues (Li, Ruiz-Primo, Giamellaro, Wills, Mason, & Lan, 2012a) which focused on the same experiment of proximity.

reported higher levels of motivation and lower level of task difficulty on the more familiar context compared to the less familiar setting.

The three studies demonstrate different sources of context familiarity but all point to the positive effect of context familiarity on students' performance. More evidence is provided in the literature to support the argument that familiar context might enhance performance. For instance, Crisp (2011) found familiar contexts that promoted the use of everyday knowledge was associated with lower item difficulty compared to unfamiliar contexts. This finding is in line with the previous study by Brownell and Stretch (1931). To investigate the influence of unfamiliar context on students' performance, they presented four sets of arithmetic problems in four contexts to 256 Year 5 students in America. The four contexts included boy scouts, soldiers' cavalry, refining oil plant, and Hindu village. Results showed that item difficulty significantly increased as context familiarity decreased.

EN_07_C

A student added 2 spoons of salt to 1 liter of water in an aquarium. She then added some brine shrimp eggs.

How would the student know if 2 spoons of salt in the water was brine shrimp's optimum condition for salinity?

A. Most of the brine shrimp eggs hatched.
B. Some of the brine shrimp eggs hatched.
C. At least one of the brine shrimp eggs hatched.
D. None of the brine shrimp eggs hatched.

EN_07_P1

A scientist made the water of a tank salty by adding 2 spoons of salt for every liter of water. She then added some crab eggs to the tank.

How would the scientist know if 2 spoons of salt per liter of water was the crab's optimum condition for salinity?

A. Most of the crab eggs hatched.
B. Some of the crab eggs hatched.
C. At least one of the crab eggs hatched.
D. None of the crab eggs hatched.

EN_07_P2

A mushroom grower wanted his mushroom barn to have the right humidity. He decided to keep the air at 70% humidity.

How would the grower know if 70% humidity was the optimum condition for the mushroom?

A. Most of his mushrooms grew.
B. Some of the mushrooms grew.
C. At least one of the mushrooms grew.
D. None of the mushrooms grew.

Figure 3.1 Example of an item bundle developed based on the *proximity* of the items to the enacted curriculum: Close (C), Proximal (P1), and Proximal 2 (P2) (Li et al., 2012a).

Shared general information	
Climate change can directly affect human health by increasing the number of extreme heat waves. According to a study conducted by a research centre at University of Chicago, increases in temperature may lead to more extreme heat waves during summer. Heat waves could occur more frequently, become more intense and last longer, even though they are in frequent events that differ in nature and consequences now.	
The US context During the 1990s, Chicago experienced several severe heat waves. In July 1995, a heat wave caused 485 heat-related deaths and 739 excess deaths when the temperature was over 98 °F for 4 days. After consecutive heat events, Chicago implemented a Heat Health Watch/Warning System. The warning system was implemented in other US cities, such as Cincinnati, New Orleans and St. Louis.	Korean context During the 1990s, Seoul experienced several severe heat waves. In July 1994, a heat wave caused 254 heat-related deaths and 532 excess deaths when the temperature was over 38 °C for 14 days. After consecutive heat events, Seoul implemented a Heat Health Watch/Warning System. The warning system was implemented in other Korean cities, such as Busan, Chunan and Gwangju.

Figure 3.2 Shared general information and example cases in two different contexts (Song & Bruning, 2016).

The positive impact of context familiarity on students' performance could be largely explained by the schema theory. According to the schema theory, schema represent a knowledge structure that guides the retrieval and encoding of information and allows for the assimilation of ambiguous passages based on readers' background and life experiences (Anderson et al., 1977). Setting assessment items in familiar contexts are more likely to activate relevant schema, which helps students encode information more meaningfully and efficiently (Song & Bruning, 2016). With appropriate schema being evoked by familiar contexts, it is believed that students could benefit in two ways, particularly when they are taking assessments. First, familiar contexts facilitate comprehension of assessment items. According to Anderson et al. (1977), high-level schema provide the interpretive framework for comprehending texts. It helps students better decode and make sense of information in the questions by making connections to what they are reading to what they know, and allows them to infer or guess meanings of ambiguous messages

from the text. Second, it helps reduce cognitive loads. Specifically, it facilitates memory searches by "providing frameworks for structuring a text and by offering guides to the information that will be recalled" (Song & Bruning, 2016, p. 693) and by filtering out unimportant or irrelevant information (Lee, 2011; Song & Bruning, 2016). The work by Song and Bruning (2016) provided a good explanation for how context familiarity reduced cognitive load as follows:

> *When students access information within familiar contexts – in the case of our participants, contexts related to conditions in US cities and states that already were schematically organized in their long-term memory, working memory limits are less likely to be reached. Presumably, the familiar geographical information could be comprehended with little or no effort, with automatic schema-driven processing and contextual information freeing working memory capacity* (Song & Bruning, 2016, p. 708).

Issues Regarding the Use of Familiar Contexts. Despite the wide discussion on the role of context familiarity on students' performance, the results reported in the literature were variable. In fact, numerous studies find that setting test items in familiar contexts is associated with neutral or negative test performance. Potential reasons are explored by empirical studies.

A major concern regarding the use of familiar contexts in assessment items is that familiar contexts may drive students' attention away from what is intended to be assessed and undermine students' performance. In the study by Boaler (1994), fifty students were given two sets of items, which measured the same mathematical content but were embedded in different contexts, including two abstract and four real-world contexts (football season, planting plants, cutting wood, and a fashion workshop). Results found that 16 females underachieved in contexts that may appear to be more familiar to them (fashion shop) compared to their performance on the abstract items or items involving football or wood cutting. Boaler (1994) speculated that the underachievement on the fashion item was because the familiar context distracted students from the intended constructs. She further explained that female students showed great involvement

with the context and tended to over-focus on the details presented in the context, as some of them discussed the nature of the job and some of them considered the order that the jobs would have to be encountered in, etc.

Aligned with the work by Boaler (1994), Almuna Salgado (2010) reported similar findings that students tended to bring in personal information into argument rather than using a mathematical argument when solving mathematical problems. In the study, Almuna Salgado (2010) compared the performance of 30 Year 10 students on two sets of items—four PISA mathematical items and the variation of the four PISA items with more familiar contexts that were closely related to students' everyday life (e.g., train ticket price). A qualitative analysis of interviews revealed students' tendency to use their personal experiences to judge and guide their solutions. For example, students considered a familiar context as real using Melbourne metropolitan ticket prices to judge their answers. Personal experiences may assist in problem solving when it aligns with the context of the assessment item. However, in cases when the information presented in the item context is not consistent with or is an alteration of the factual knowledge that students may obtain from the real world, relying on personal experience rather than intended constructs may mislead students and result in low test performance.

In addition to the major concern described above, the empirical research also documented other challenges associated with the use of familiar contexts and how it impacts students' performance. In a follow up study, Almuna Salgado and Stacey (2014) found that when students were familiar with a context, they tended not to give a very detailed answer, which may cause them to lose points especially when the item requires constructed responses. One example given by Almuna Salgado and Stacey (2014) is that students did not explain that their answers were based on proportional reasoning when responding to a PISA mathematical item and therefore

they did not get full credit. Almuna Salgado and Stacey (2014) speculated that students did not provide detailed answer because the context was so familiar that students assumed everyone already knew the arguments.

Besides, Ahmed and Pollitt (2007) note that setting items in familiar contexts may result in additional cognitive loads. When students are presented with a familiar context, it is likely that a great amount of information related to the context are activated in their long-term memory and become accessible. Students have to select relevant information while compressing the irrelevant ones, which leads to greater mental efforts and may impact their performance during the test.

Furthermore, the study by Anderson et al. (1977) not only found students performed better on items that were set in a context related to their educational background, but also showed that most students selected only one distinct interpretation for each question. Moreover, 80% of participants reported they were unaware of the other perspective while reading the passage. This indicates that familiar contexts may direct an individual to perceive the question in a certain way without even considering alternative interpretations. Based on the findings, it follows that setting items in familiar contexts may constrain students' divergent thinking. This may pose negative influences on students' performance when they are dealing with tasks or items that have multiple correct answers (e.g., multiple-answer multiple choices) or require more than one way of reasoning or interpretation (e.g., constructed responses requiring at least two reasons which could explain the observed phenomenon.)

Conclusion. In summary, an overview of the literature above reveals three research gaps that this study attempts to address. First, previous literature demonstrates two major sources of familiarity. One is students' personal life such as their cultural background (Song & Bruning, 2016) and daily events (Ahmed & Pollitt, 2000). The other source is related to curriculum and

classroom practices (Li et al., 2012a; Li et al., 2012b). Most of studies focus on only one source of familiarity, which may result in an inadequate measure of students' familiarity with a context. For instance, a student may indicate he/she did not have daily life experience with a certain context, which leads researchers to believe he/she is unfamiliar with the context, whereas the student may have been exposed to the context from other channels such as in his/her physics classes. This study fills the gap by taking into account both sources of context familiarity.

Second, familiarity with a context is closely related to individual experiences and perceptions. However, the general experiment design used in most of studies is dividing students into multiple groups. Based on some group features (e.g., gender, nationality, etc.), researchers decide whether each group is familiar with the context or to what extent each group is familiar with the context. One drawback of such an experiment design is that it neglects the within-group variability. Even in the same experiment group, some students may be more familiar with the context than others. Ignoring individual differences may lead to inaccurate impressions on how context familiarity impacts group performance, which further affects the curriculum development and instructional design for different groups of students. In this study, we measure context familiarity based on students' self-reported responses to eight survey research questions, rather than pre-defining the degree of context familiarity for different student groups from researcher's view.

Third, most studies focused on the effect of context familiarity on either student scores or the item difficulty index. As prior work has proved context familiarity could impact both outcomes, in this study we advocate for the need to take into account both students' performance and item statistics in order to provide a more comprehensive picture of the effect of context familiarity.

3.2.2 Cognitive Diagnostic Model

Cognitive diagnostic models (CDMs) are latent variable models developed primarily for assessing students' mastery and non-mastery of a set of fine-grained skills (de la Torre, 2011). The key idea of a CDM is that it takes the experts' cognitive model of the domain knowledge and turns it into a probability model that classifies examines into latent classes based on their observed response patterns and a user-defined skill matrix (Javidanmehr & Sarab, 2017; Sinharay & Almond, 2007). Previous research has proposed a variety of CDMs based on different assumptions about how attributes influence students' responses in assessments. A *conjunctive* model assumes students have to possess all required attributes in order to answer the item correctly. The deterministic inputs, noisy, "and" gate (DINA; Haertel, 1989) model is one of the most widely used conjunctive models, which assigns the highest probability of a correct response to examinees that possess all of the required attributes (Ma, Iaconangelo, & de la Torre, 2016). A *disjunctive* model, on the other hand, assumes that students only need to possess a subset of attribute to successfully solve an item. For instance, the deterministic inputs, noisy, ''or'' gate (DINO; Tempen & Henson, 2006) model assigns the highest probably of answering correctly to examines who possess at least one of the required attributes (Ma et al., 2016). There are also more generalized models such as the general diagnostic model (GDM; von Davier, 2008) and the generalized deterministic inputs, noisy "and" gate model (G-DINA; de la Torre, 2011).

A large number of studies have concentrated on the application of CDMs to measure students' competency in science and mathematics large-scale assessments (e.g., Yamaguchi & Okada, 2018). The increasing popularity of CDM in science and mathematics assessments can be attributed to its ability to offer diagnostic information. More specifically, it provides information

about students' strength and weakness with respect to the set of knowledge and skills assessed, which are commonly referred as *attributes* (Cui, Gierl, & Chang, 2012). One may argue that traditional assessments could also provide such information. For instance, the PISA 2015 Assessment and Analytical Framework listed three competencies assessed in the PISA science test: explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically. However, such competencies are defined at coarse grain size, or in another words, "the level of specificity with which one would like to make statements about respondents" (Rupp, Templin & Henson, 2010, p. 52), which turn out to be too broad in the scope. The CDM fills in the gap by focusing on micro-level skills, albeit attributes, that are specified at fine grain size, which helps align assessment results with the goals of curriculum and instruction (Lim, 2015; Rupp et al., 2010; Templin & Bradshaw, 2013) and provide diagnostic information of students' learning that could directly inform instructional practices.

In CDM, the link between test items and attributes are established through a *Q-matrix*, which is an index matrix specifying the attributes required for each item. The Q-matrix is usually grounded on the solid theories of the latent constructs and needs to be specified based on substantive knowledge of content experts and cognitive psychologists (Chen, de la Torre, & Zhang, 2013; Jang, 2009). Importantly the development of attributes and the specification of Q-matrix should be guided by assessment purpose and use. For example, if the intended use is to obtain some diagnostic information about students' mastery status of a particular topic in a unit, attributes then need to be defined in a relatively small scope in order to target the most essential concepts. The number of attributes may also be constrained given the number of students and items available. On the contrary, a large-scale assessment is able to assess more attributes across multiple units but at a coarse level and may have a more complex Q-matrix design (e.g., a

hierarchical structure of attributes) so that the test result could provide rich diagnostic information about students' weakness and strength for those key sub-domains and offer accurate statistical inference for the diverse student sample.

G-DINA and Reduced Models. The G-DINA model (de la Torre, 2011) is a generalization of the DINA model, which is one of the simplest and most interpretable CDMs for dichotomously scored test items. As the DINA model assumes attribute vectors in the same group have an identical probability of answering an item correctly, the G-DINA relaxes this assumption by partitioning latent classes into 2^{K_j} (K is the number of attributes) groups and allowing the probability of correct answers to vary across groups.

The G-DINA model is based on a J x Q matrix. Consider a test with a J x K Q-matrix, the element in row j and column k of the Q-matrix, q_{jk}, is equal to 1 if the kth attribute is required to answer item j correctly, and zero otherwise. As not all items measure all attributes, we denote K_j^* as the number of required attributes for item j, yielding $2^{K_j^*}$ attribute profiles. For example, an item measuring 2 attributes will have 2^2 profiles: (00), (10), (01), (11), with 1 and 0 representing the presence and absence of the first and second attribute, respectively. Each profile can be considered a latent group for item j, represented by an attribute vector $\alpha_{lj}^* = \left(\alpha_{lj1}, \ldots, \alpha_{ljK_j^*}\right)'$. Let $P\left(X_j = 1 | \alpha_{ij}^*\right)$ denotes the probability of correct response for item j given α_{ij}^*, the reduced attribute vector consisting of required attribute for item j. The item response function (IRF) of the G-DINA model can be expressed as

$$P\left(X_j = 1|\alpha_{ij}^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (3.1)$$

where δ_{j0} is the intercept for item *j*, which represents the baseline probability of answering item *j* correctly when none of the required attributes is present; δ_{jk} is the main effect due to the *k*th attribute α_k; $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$; and, similarly, $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$.

Several commonly used CDMs could be derived from the G-DINA model by setting appropriate constraints and link functions (de la Torre, 2011; Ma et al., 2016). Five reduced (constrained) models were represented below as constrained cases of the G-DINA model.

The DINA model can be obtained by setting all the parameters, except δ_{j0} and $\delta_{j12\ldots K_j^*}$, to zero:

$$P\left(X_j = 1|\alpha_{ij}^*\right) = \delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (3.2)$$

As shown in the model formulation, the DINA model has two parameters per item: g_j (δ_{j0}) is the probability of *guessing* an item correctly when examines lack one or more prescribed attributes for item *j*, and 1-s_j ($\delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}$) is the probability that examines who have all the required attributes not slip and answer the item correctly. Based on 1-s_j, we could easily calculate the probability of *slipping* and getting the item wrong ($s_j = 1 - P\left(X_j = 1|\alpha_{ij}^*\right)$).

Similar to the DINA model, the deterministic input, noisy, "or" gate (DINO) model can be derived from the G-DINA model by constraining the magnitudes of the main and interaction effects to be identical to each other:

$$\delta_{jk} = -\delta_{jk'k''} = \cdots = (-1)^{K_j^*+1}\delta_{j12\ldots K_j^*} \tag{3.3}$$

where $k = 1,\ldots, K_j^*$, $k' = 1,\ldots,K_j^*-1$, and $k'' > k',\ldots, K_j^*$. Specifically, the IRF of DINO is given by

$$P\left(X_j = 1|\alpha_{ij}^*\right) = \delta_{j0} + \delta_{jk}\alpha_{lk} \tag{3.4}$$

The DINO model also has two parameters per item. Guessing (δ_{j0}) for the DINO model is the probability of a correct response when none of the required attributes is mastered by examinees, while slip (1- $P\left(X_j = 1|\alpha_{ij}^*\right)$) is the probability of a wrong response for examinees who have at least one of the attributes.

Three additional models can be derived from the G-DINA model. By setting the interaction effects to be zero, the additive cognitive diagnostic model (ACDM) could be obtained:

$$P\left(X_j = 1|\alpha_{ij}^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{lk} \tag{3.5}$$

The linear logistic model (LLM) is the logit link G-DINA model without the interaction terms, and its IRF is given by

$$logit[P\left(X_j = 1|\alpha_{ij}^*\right)] = \delta_{j0} + \sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{lk} \tag{3.6}$$

The reduced reparametrized unified model (RRUM) is the log-link G-DINA model without interaction terms and it is formulated as

$$log[P(X_j = 1|\alpha_{ij}^*)] = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} \tag{3.7}$$

ACDM, LLM and RRUM are all additive models which assume that mastering attribute α_{lk} will increase the probability of success on item *j* and the effect is independent of the effects of other attributes. All three models have K_j^*+1 parameters for each item.

In this study, the *guessing* probability for the all models were calculated as the probability of a correct response in the absence of all required attributes (which is equivalent to the intercept δ_{j0}), and the *slip* probability is calculated as the probability of answering an item wrong when all required attributes are present.

***Model Comparison and Selection*.** CDM requires the Q-matrix to be well-specified, whereas in lots of situations the nature and the relations of attributes turns out to be ambiguous. In this case, it is suggested to start with fitting a saturated CDM. According to Chen and Chen (2016b), a saturated CDM takes into account all possible item parameters, including interactions of attributes at the item level. In contrast, a reduced CDM only allows for a few patterns of attribute interactions. Theoretically, saturated models always yield better model-data fit than reduced models given their complex parameterization. However, it usually requires larger sample sizes to be estimated precisely and the results may be hard to interpret (Chen et al., 2013). In contrast, reduced models may be preferred for three reasons. First, reduced CDMs requires smaller sample size and are easier to interpret. Second, appropriate reduced models can provide better classification rates than saturated models, especially when the sample size is

small. Last, according the parsimony principle, the simpler model is preferred when it does not fit the data significantly worse than the saturated model (Ma et al., 2016).

The comparison of different models could be performed at the test level and the item level using various model fit indices. There are two major groups of model fit indices: relative fit, whose value has to be interpreted with reference to other competing models, and absolute fit, which checks whether the model at hand fits the data adequately. There two groups of model fit measures are well studied by Chen et al. (2013). In their study, they examined the sensitivity of three relative fit indices — Akaike's information criterion (AIC), Bayesian information criterion (BIC), and –2log likelihood (–2LL). For each of these three statistics, the model with smallest value is considered most appropriate among a set of competing models. All three statistics could be computed as a function of the maximum likelihood (ML). In addition, they reported results on three absolute fit indices. They were residuals based on the proportion correct of individual items (ρ), Fisher-transformed correlation of item pairs (r), and the log-odds ratio of item pairs (l). The three absolute fit indices are based on the differences between the observed responses and the predicted responses generated from the fitted model. To determine whether the residuals are significantly different from zero, one needs to calculate standard errors to obtain z-score for each measure. It should be noted that all three statistics could be used to evaluate individual item fit. To extend the use to the test-level model comparison, we use the maximum z-score of each statistic. For all three absolute measures, the value is expected to be close to zero if the model fits the data; otherwise, as Chen et al. (2013) indicated, rejection of any z-score implies that the model does not fit at least one item or item pairs adequately.

According to Chen et al. (2013), BIC was the most reliable for detecting misspecifications of the CDM, the Q-matrix or both among three relative fit indices. AIC

performed pretty similarly to BIC with a relatively lower sensitivity to misspecification to all three conditions, while the performance of -2LL seemed to be unreliable in most cases. For absolute fit indices, the proportion correct transformed correlations and log-odds ratios had similar performance and were sensitive to different misspecifications in most conditions, while the proportion correct failed to detect the misspecification consistently in most cases.

Besides the model fit indices discussed above, the Wald test is one of the methods that allow for the model comparison at the item level. de la Torre (2011) suggested to use the Wald test to examine whether reduced models will could be used in place of the saturated models without significant loss in model-data fit. To perform the Wald test, one needs to set up R_{jp}, a $\left(2^{K_j^*} - p\right)$ x $2^{K_j^*}$ matrix of restrictions for item j and the reduced model p, where p represents the number of parameters of the reduced model and K_j^* is the number of required attributes for item j. Specific constraints need to be set to each of the models. Then the Wald statistics W is computed as

$$W = \left[R_{jp} \times f\left(P_j\right)\right]'\left\{R_{jp} \times \text{Var}\left[f\left(P_j\right)\right] \times R_{jp}'\right\}^{-1}\left[R_{jp} \times f(P_j)\right] \quad (3.8)$$

and is asymptotically $\chi^2_{2^{K_j^*}-p}$ under the null hypothesis that R_{jp} x $f(P_j)$ = 0. de la Torre (2011) also noted that the implementation of the Wald test did not require the reduced model to be estimated as long as $f(P_j)$, Var$[f(P_j)]$ and R_{jp} were known. This method has been proved to maintain relatively accurate Type I error with large smaller sizes and small number of parameters (de la Torre & Lee, 2013). It is also found that the Wald test has excellent power to detect the true model at reasonable significant levels (e.g., 0.05) when the true model is DINA, DINO or ACDM even for relatively small sample sizes.

Outputs of G-DINA. After the best set of CDMs are fitted to data, it is important to turn those sophisticated statistics from the model into something interpretable and meaningful. According to Chen and Chen (2016b), outputs of the G-DINA model includes five major components: (1) the relative and absolute model fit statistics indicating the fitness of the model to the data; (2) the attribute prevalence, which shows the whole sample's mastery probability of each attribute; (3) the latent classes based on the attributes and their posterior probabilities; (4) the individual examinee's mastery probability of each attribute, based on which the G-DINA model provides individual *attribute profiles* which record the mastered (represented as 1) and non-mastered (represented as 0) attributes for each student; and (5) the estimates of item parameters, such as *guessing* and *slip*. The results of the current study touch upon on all five outputs, of which the last two will be further explored in terms of their relationship with the effect of context familiarity.

3.3 METHOD

3.3.1 Data

This study collects 1,631 students' responses to a physics assessment, of which 866 were middle-school students and 765 were high-school students. There are roughly equal number of males (50%) and females (48%), with 2% of students did not disclose their gender.

The physics assessment is designed to assess students' understanding of Forces and Interactions at the middle school level. Specifically, the content of the assessment is carefully developed align with Grade 8 physics standards and the reading load is controlled to be accessible for students at 8^{th} grade.

At the end of the assessment, students were asked to respond to eight survey questions, at the last page of the test booklet, regarding their familiarity with the object(s) of the context in

assessment items (details about survey questions are provided in the later section). Only 1,478 students responded to the survey questions. As survey questions are important indicators of students' familiarity with item contexts, our data analysis excluded students who did not respond to the survey questions, leading to a student sample size of 1478 for this study[6]. Students' responses to assessment items were dichotomized with 1 being correct and 0 being incorrect. Similarly, students' responses to survey questions were dummy coded based on whether or not students indicated they had experiences with the object(s) of interest described in the item context.

3.3.2 Instrument

This instrument was based on a test that we developed in another study which examined the effects of context richness on students' performance (Dong, Li, Ruiz-Primo, Zhai, & Minstrell, 2018). Based on the feedback from teachers and students and the evolvement of research focus, the current instrument shows substantial changes in item design.

Fundamental Ideas. The instrument used in this study focuses on a critical *Next Generation Science Standards* (NGSS) disciplinary core idea, Forces and Interactions, at the middle school grade band. In collaboration with experts in physics education and secondary physics teachers, the core idea of Force and Interactions was unpacked into smaller, fundamental understandings, which we name as *Fundamental Ideas*. The original list of fundamental ideas went through a rigorous, iterative process of revisions based on inputs from domain experts

[6] We calculated the proportion of correct response for each test item for the group of 1,478 students and for the group comprised of 153 students who were excluded from the analysis. A two-sample *t*-test was conducted between two sets of proportion of correct response for the two groups. Results of the *t*-test shows there is no significant difference in the proportion of correct response of test items between the two groups ($t(68) = 0.24$, $p = 0.815$), which indicates that the performance of those 153 students excluded from the analysis did not show patterns that are significantly different from the larger group in terms of responses.

including physics researchers, middle and high school physics teachers, and measurement experts. The final list of fundamental ideas contains 29 fundamental ideas organized into seven groups. In this study we target eight out of 29 fundamental ideas. The target fundamental ideas are provided in Table 3.1.

Table 3.1
Definitions of eight fundamental ideas

A1	Each force has a magnitude and a direction.
A2	Horizontal forces' relative sizes and directions can be added. If horizontal forces in each direction are added to obtain a sum of forces and the sums in opposite directions are not equal, then there is a net force (aka unbalanced force) in the direction of the larger sum and the magnitude is the difference between the horizontal sums.
A3	If at rest, the object may have zero or two or more forces acting on it, but the forces add to give zero net force.
A4	If an object is moving at a constant velocity, then the net force is zero.
A5	If the net force is positive (in the direction of motion), then the object will speed up.
A6	If the net force is negative (in the backward direction from the motion), then the object will slow down in the forward direction until zero or speed up in the backward direction.
A7	Two objects in contact exert (static/kinetic) frictional force on each other and act parallel to the contact surface.
A8	A greater force will cause a greater change in motion for the same mass (mass is kept constant).

Note. Motion in the fundamental ideas above is assumed to be in the forward direction.

Clusters of Items. This instrument contains four *clusters* of items. A cluster resembles a testlet, a commonly used structure to organize a bundle of items under a shared stimulus or context (Jiao, Wang, & Hentia, 2013; Wainer, Bradlow, & Wang, 2007). Compared to a testlet, a cluster presents a more complex nesting structure of items. According to Ruiz-Primo and her colleagues (Ruiz-Primo, Li, Minstrell, Kanopka, Hernandez, Dong, & Zhai, 2019), a *cluster* is consisted of a general scenario (written as a paragraph) followed by four *sub-testlets* corresponding to that scenario. The four sub-testlets intend to assess a student's understanding of

motion in four conditions: constant velocity, acceleration (speeding up), deceleration (slowing down), and at rest. Each sub-testlet contains a two-tiered set of items (Treagust & Haslam, 1986): the first item asks students to describe the motion of the objects given a scenario, and the second item focuses on explaining the forces that are involved or associated with the state or the change of motion in item 1. Figure 3.3 shows a prototype of a cluster proposed by Ruiz-Primo et al. (2019).

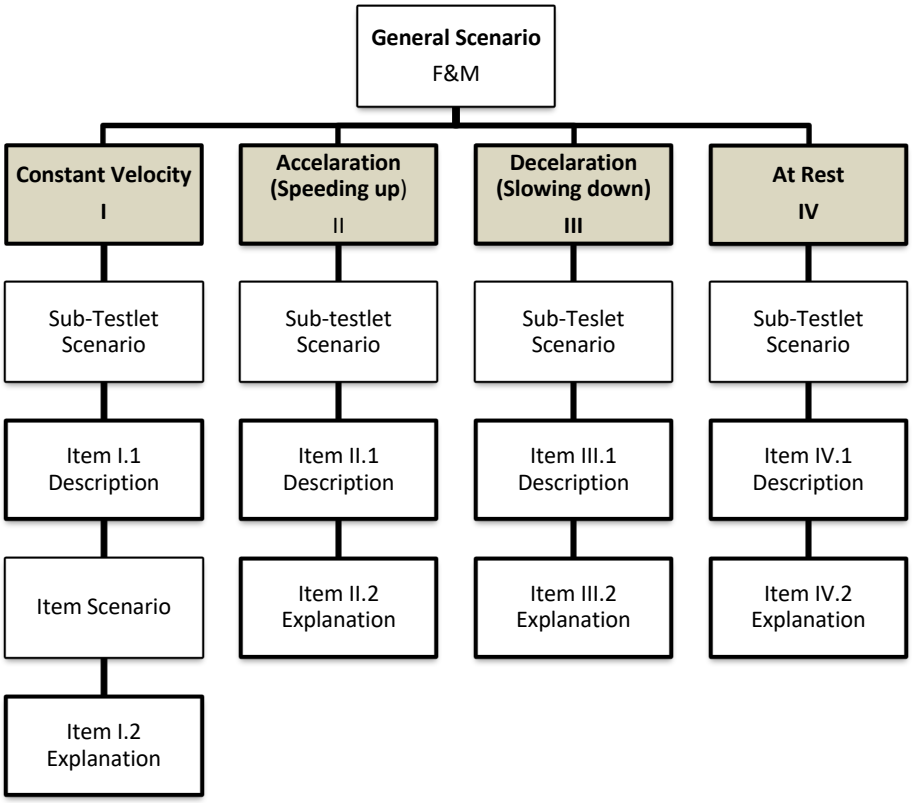


Figure 3.3 Prototype of a cluster (Ruiz-Primo et al., 2019).

In this study, we developed four clusters with different contexts, each tapping all these four conditions (constant velocity, acceleration, deceleration, and at rest). Two of the clusters describe scenarios with vehicles: (1) a school bus on a road and (2) a dogsled in snow. The other two clusters describe scenarios with students moving objects in a school setting: (1) pushing a box on a gym with different types of floors and (2) pulling a cart of books on a library with

different types of floors. Clusters were developed as parallel as possible on all characteristics (e.g., assessed fundamental ideas and reading load) but the object(s) of interest and the scenario (Ruiz-Primo et al., 2019). To provide a concrete example of a cluster, we present the general context of the box cluster and one of the sub-testlets tapping acceleration (see Figure 3.4).

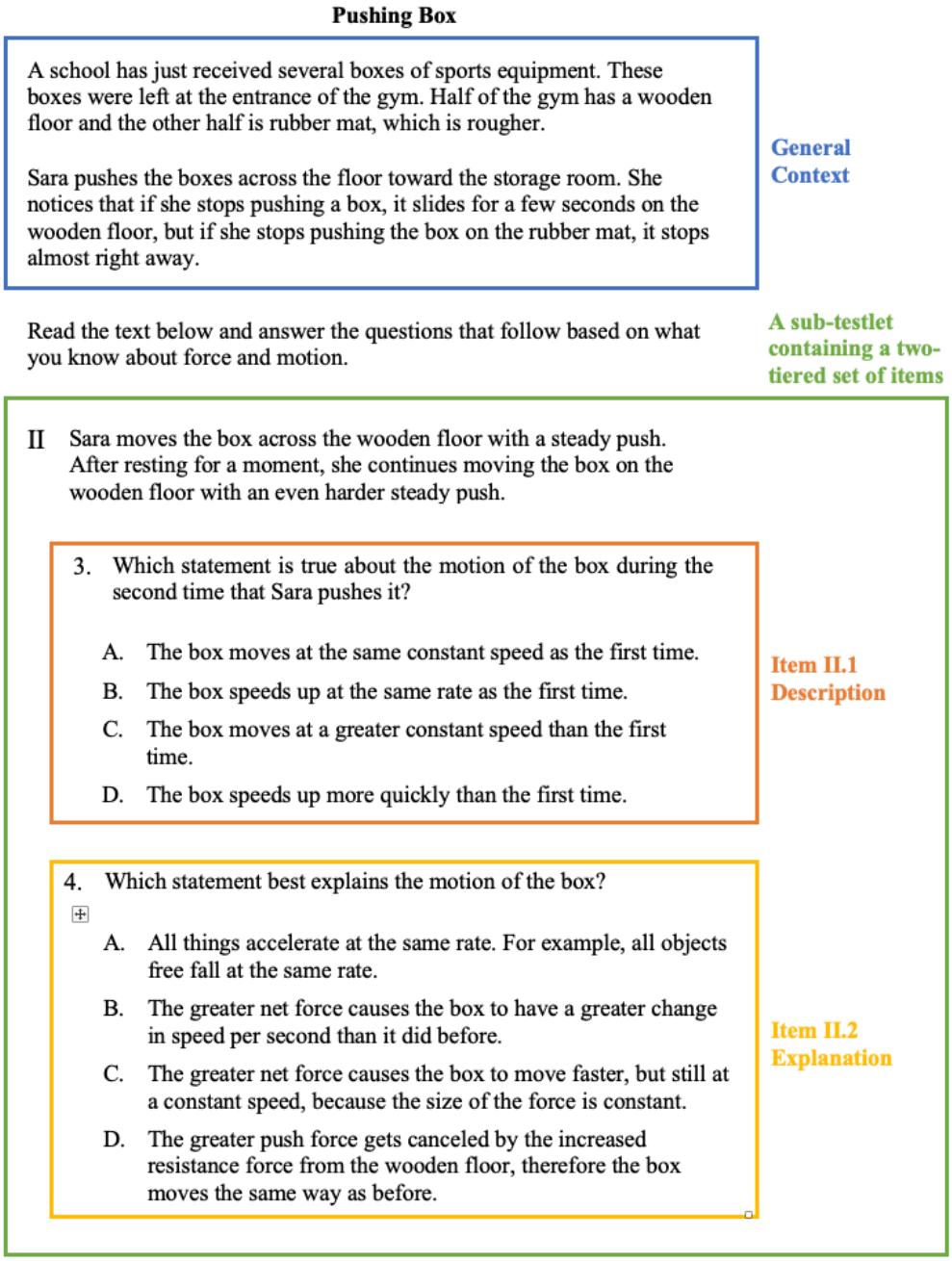


Figure 3.4 The box cluster with the general context and a sub-testlet tapping acceleration

Booklet Design. A total of 32 items (four clusters × four sub-testlet per cluster × two items per sub-testlet) were evenly distributed into two booklets. To link the performance of the two groups of students who received different booklets, three common items were added to each of the booklet at the exact same place. The three common items were selected from the publicly available items from two research projects (AAAS at assessment.aaas.org and Diagnoser at www.diagnoser.com) with known item statistics. In total each booklet contains 19 items. The two booklets were randomly distributed to students within a class, resulting in a sample size of 747 and 731 for booklet 1 and 2, respectively. The two groups' performance were equated using the simultaneous calibration approach (de Ayala, 2009).

Familiarity with Contexts. An eight-question survey was attached at the end of the test to investigate students' prior experiences with the object(s) of interest described in the general context [7]of each cluster. For each cluster, students needed to respond to two survey questions. For example, the two survey questions for the box cluster are:

1. Have you ever pushed a box on a wooden floor?
2. Did you discuss and/or use an example of a person pushing a box in your physics class?

The survey questions were almost identical for each cluster except for the change in the object(s) of interest. The two survey questions served as proxy of the two sources of context familiarity: daily life experience and classroom experience. It should be noted that the survey questions correspond to the general context of the cluster rather than individual item contexts. Therefore, if a student answered yes to survey question 1 and 2 for a cluster, his responses will be used for all items in that cluster.

7 The object(s) of interest may also appear at the sub-testlet level and individual item level. Regardless of the level of cluster it appears, the object of interest remains the same as long as the items are in the same cluster.

3.3.3 Q-matrix

Using the eight fundamental ideas as attributes, a 16 x 8 (16 items x 8 attributes) Q-matrix was created by physics experts and secondary physics teachers to map each item with its required fundamental ideas. The original Q-matrix was tested in a field testing of 316 middle-school students using the G-DINA model. Based on the Q-matrix validation results suggested by the G-DINA model, the Q-matrix was reviewed and revised by domain experts. As the four clusters in the present study are designed to be as parallel as possible on the intended constructs, the Q-matrix for each cluster is identical. For brevity, we present the final Q-matrix for the box cluster in Table 3.2.

Table 3.2
The Q-matrix for the box cluster

Item ID	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
Box I.1	1	1	0	1	0	0	0	0
Box II.1	0	1	0	0	0	1	0	1
Box III.1	0	0	0	0	1	0	1	1
Box IV.1	1	1	1	0	1	0	0	0

3.3.4 Analysis

The analysis procedure contains three major steps. The first and most essential step is to apply an appropriate CDM to the dataset. For this work we use the G-DINA model and reduced models. The G-DINA model family produces two kinds of parameters: person parameters and item parameters. The person parameters represent each individual student's attribute pattern. The item parameters including guessing and slip for each item. After the CDM is selected and parameters are estimated, regressions are applied to model the effect of context familiarity on

person parameters and item parameters, respectively. The detailed analysis procedure is described as follows.

Selection of CDM. To select an appropriate CDM, a G-DINA model was fitted to the data as the first step using the pre-defined Q-matrix. This model serves as a baseline for understanding how the items behave overall and what may be some potential issues with the model and the Q-matrix. Next, five reduced models were compared to the G-DINA model using the Wald test (Ma et al., 2016). The reduced model with the largest non-significant *p*-value was selected for each item. If none of the reduced models was found to be significant, the G-DINA model was retained. It should be noted that the G-DINA model also provides suggestions on how to revise the Q-matrix to improve the model fit. However, the Q-matrix validation suggested by the G-DINA model was based on empirical data. For the present study we decide to use our pre-defined Q-matrix without any revisions. This is because the development of our Q-matrix involves extensive efforts from domain experts and were carefully crafted to target attributes of our interest, which, we believe, provides valuable information about students' learning from a theoretical perspective.

The model selection was conducted using the relative and absolute model fit indices introduced by Chen et al. (2013). The G-DINA models and reduced models were compared based on AIC, BIC and Likelihood Ratio Test (LRT). Another commonly used relative model fit index -2loglikelihood (-2LL) was not reported in this study as Chen et al. (2013) found -2LL was not reliable in detecting misspecifications of the model and the Q-matrix. Next, the final set of models selected was evaluated using the three absolute model fit indices: proportion correct, transformed correlation, and log-odds ratio.

For the final set of models, the item discrimination indices (de la Torre, 2008) was calculated for each item to evaluate potential misspecifications of the Q-matrix. The discrimination indices were calculated as the difference between the probability of a correct response without mastering any required attributes and the probability of a correct response with all required attributes being mastered (de la Torre et al., 2015). Following the suggestions by de la Torre et al. (2015), a cut-off value of 0.4 was used to consider an item as discriminating.

All analysis was conducted using the R package *GDINA* (Ma, de la Torre, & Sorrel, 2018). Marginal maximum likelihood method with Expectation-Maximization (MMLE/EM) algorithm was used for item parameter estimation.

Regressions with Person Parameters as Dependent Variables. After the appropriate CDMs are selected, an individual attribute profile was calculated for each student using the Expected A Posteriori (EAP) estimation. The estimated individual attribute profile contains eight dummy numbers representing the mastery status of eight attributes (1 for mastering and 0 for not mastering). For instance, a student who is estimated to master A5 and A8 will have an attribute profile of "00001001". Taking each of the eight attributes as a dependent variable, the attribute pattern was regressed on students' responses to eight survey questions (all responses were dummy coded). For example, the model representation for predicting the probability of mastering attribute A1 is presented as follows:

$$P(A1 = 1) = \frac{1}{1 + e^{-Z}} \tag{3.9}$$

where

$$Z = \beta_0 + \beta_1(bus\_survey1) + \beta_2(bus\_survey2) + \beta_3(sled\_survey1) + \beta_4(sled\_survey2)$$
$$+\beta_5(box\_survey1) + \beta_6(box\_survey2) + \beta_7(cart\_survey1) + \beta_8(cart\_survey2)$$

The *p*-value was adjusted based on Bonferroni method to account for the inflated type I error due to multiple regression models.

To check the accuracy of person parameter estimation, a simulation study was conducted. Firstly, we simulated students' responses based on the estimated guessing and slip parameters using the *simGDINA* command in the *GDINA* package. Next, the classification rates (proportion of correctly classified attributes) were calculated for two sets of attribute profiles (simulated vs. true estimated attribute profiles) using the *ClassRate* command in the *GDINA* package. The classification rates (CR) was calculated at two levels: an overall CR considering attribute profiles of all attributes; and CRs for each individual attribute.

Regressions with Item Parameters as Dependent Variables. At the item level, we are interested in the relationship between context familiarity and two item parameters—guessing and slip. As items rather than students become the unit of analysis at this step, we aggregated students' responses to survey questions and created two item-level indicators to measure how familiar students were in general to a certain cluster context. More specifically, we calculated the mean of students' responses for each survey question, which indicated the proportion of students who responded "yes" to each survey question. Using 0.4 as a threshold, survey questions with a mean above 0.4 were coded as 1, meaning that the majority of students had either daily life or classroom experience with the cluster context. In other words, the cluster context was considered familiar to most of students in terms of one or both sources. Otherwise survey questions were coded 0, which indicated most of students did not have daily life and/or classroom experience (and thus were unfamiliar) with the cluster context. The cut-off value of 0.4 was selected as it was close to the middle point of 0.5 (none of survey questions reported a mean of 0.5) and, more

importantly, it divided items into two groups that were relatively balanced. Table 3.3 shows the two item-level familiarity indicators for each item.

Table 3.3
Indicators of context familiarity for each item by aggregating students' responses to survey questions

Item ID	Daily Life Experience	Classroom Experience	Item ID	Daily Life Experience	Classroom Experience
Bus I.1	1	0	Box I.1	1	1
Bus II.1	1	0	Box II.1	1	1
Bus III.1	1	0	Box III.1	1	1
Bus IV.1	1	0	Box IV.1	1	1
Sled I.1	0	0	Cart I.1	1	1
Sled II.1	0	0	Cart II.1	1	1
Sled III.1	0	0	Cart III.1	1	1
Sled IV.1	0	0	Cart IV.1	1	1

Note. 1 indicates the majority (more than 40%) of students had either daily life or classroom experience (and therefore were familiar) with the cluster context in which the item is situated; 0 means the cluster context of the item was unfamiliar to most of students as less than 40% of students did not have daily life or classroom experience with the cluster context.

As seen in Table 3.3, the box and cart contexts were most familiar to students as the majority of students had experiences with the object(s) of interest both in daily life and in classrooms. The two contexts were followed by the bus context, as most of students indicated they had experiences with the bus in daily life. The sled context was the least familiar to the majority of students.

Next, the two familiarity indicators were used as independent variables to predict guessing and slip parameters, respectively[8]. Again, the statistical significance was adjusted based on Bonferroni method. In addition, we calculated Cohen's f^2 (Cohen, 1988) as the effect size

[8] We performed the regression analysis and the simulation study using continuous (mean of students' response for each survey question) and dichotomized familiarity indicators as independent variables (IV), respectively. The results from the two models were consistent. In this study, we report the results based on the model using the dummy familiarity indicators as IV as the dichotomized predictors provide a more straightforward interpretation of whether or not the item is familiar to the majority of students.

index. According to Cohen (1988), a f^2 value of 0.02, 0.15, and 0.35 is considered small, medium and large effect size.

To assess the robustness of the results, a simulation study was conducted to assess the parameter recovery of regression models. The simulation contains 30[9] replications. For each replication, a total of 1,478 student responses were simulated on the basis of the true Q-matrix and the final CDMs using the *simGDINA* command in the *GDINA* R package. The simulated responses then went through the three-step analysis procedure described above. Bias, Relative Bias and Mean Squared Error (MSE) were used to evaluate the difference between the true regression coefficients and the simulated regression coefficients. For all three criteria, a small number close to zero indicates a better parameter recovery.

3.4 RESULTS

This section is organized into three subsections in response to the three steps of the analysis procedure described in the Method section. The first subsection presents the selection of the final CDMs and important parameters estimated from the final CDMs. The second subsection reviews the results of logistic regressions which examine the effects of context familiarity on students' mastery pattern of eight attributes. The last subsection addresses the relationships between context familiarity and two item parameters—guessing and slip.

[9] We conducted the simulation study 15 times with increasing number of replications (10 to 150 increase by 10) each time. For each simulation, the parameter recovery was assessed via bias, relative bias and mean squared error, averaged across the replications. The result of the simulation study showed no significant change in three evaluation criteria after 30 replications. Therefore, in the study, we reported the results of 30 replications.

3.4.1 Cognitive Diagnostic Model

Model Selection. The model fit of the G-DINA model and reduced models was compared via three relative model fit statistics. Both AIC and BIC advocated for reduced models (AIC for the G-DINA and reduced models were 14780.28 and 14594.44, respectively; BIC for the G-DINA and reduced models were 16979.13 and 16295.24, respectively). Furthermore, the result of LRT showed there was no significant difference between the saturated model and simpler models (χ^2 (94) = 2.17, $p > 0.05$), indicating that the reduced models could replace the G-DINA model without significant loss in model-data fit.

The model fit of the selected reduced models was evaluated using three absolute fit indices. As shown in Table 3.4, proportion correct and transformed correlation reflected a good model-data fit based on the adjusted *p*-value associated with the maximum *z*-score of each statistic. However, the log odds ratio indicated at least one item or one item pair did not have an adequate fit with the Q-matrix. Such a potential alarm of model misfit or Q-matrix misspecification is not uncommon when retrofitting an assessment with a CDM. Cho (2016) applied three CDMs (DINA, DINO and G-DINA) to investigate the construct validity and psychometric prosperities of an instrument assessing emotion understanding (Situational Test of Emotional Understanding). Five Q-matrices were proposed to explain the item-attribute relationship and were compared for model fit using absolute and relative model fit statistics. The results showed none of the Q-matrices was supported by the absolute fit statistics. As Cho (2016) concluded, “this result may be inevitable when CDM is applied after the scale has been developed without considering the CDM attribute space” (p. 43).

Table 3.4
Absolute model fit statistics for the selected models

	mean[stats]	**max[stats]**	**max[z.stats]**	***p*-value**	**adj. *p*-value**
Proportion correct	0.00	0.01	0.49	0.623	1.000
Transformed correlation	0.02	0.07	2.87	0.004	0.231
Log odds ratio	0.34	3.33	23.18	0.000	< 0.001

Note. The *p*-value and adj. *p*-value are associated with max[z.stats]. The adj. *p*-values are based on the holm method.

To further examine the specification of the Q-matrix, we examined the discrimination index using a cut-off value of 0.4. As seen in Table 3.5, all items demonstrated the ability to discriminate students who mastered all versus none attributes. Given the model fit statistics and the fact that most of items showed acceptable discrimination indices, we believe the selected CDMs reflected a reliable diagnosis of students' understanding and fits the data well (although not the best).

Table 3.5 lists the final model selected for each item. As shown in Table 3.5, the G-DINA model were selected for only one item. Three items were fitted with the DINO models, which assigns the highest probability of answering correctly to examines who possess at least one of the required attributes (Ma et al., 2016). The rest 12 items all exhibited some additive effects on the response probability, of which one was fitted with RRUM and the remaining 11 with LLM.

Attribute Prevalence. Figure 3.5 presents the estimated probability of mastering each attribute for the whole student sample. Overall the attributes assessed in this test were relatively difficult for students as seven out of eight attributes reported a mastery probability lower than 0.50.

Attribute A1 and A2 seemed to be extremely difficult as only 10% of students out of the whole sample showed mastery of it. This indicates students are struggling with the two

constructs. A1 is about the nature of forces and A2 involves complex concepts about horizontal net force. The low probability of mastering attributes A3, A4 or A6 (39%, 29%, 37% of mastery probability, respectively) reflected students' confusion about the relationship between the net force and the motion. A3 and A4 involve predicting the net force given a certain kind of motion (constant velocity or at rest) while A6 targets the change in the motion when the net force is known. Attributes A5 and A7 were less prominent in terms of mastery probability (47% and 40% of mastery probability, respectively), suggesting students have a better understanding of the motion of speeding up and the friction compared to other concepts discussed above. The attribute A8, mastered by 59% of the examinees, was the easiest attribute. The mastery of A8 indicates a relatively good understanding of Newton's second law, especially the relationship between the change in force and the change in motion when mass is kept constant.

Table 3.5
Final CDMs with discrimination index

Item ID	**Model**	**Discrimination Index**	**Item ID**	**Model**	**Discrimination Index**
Bus I.1	DINO	0.60	Box I.1	DINO	0.56
Bus II.1	LLM	1.00	Box II.1	LLM	1.00
Bus III.1	LLM	0.41	Box III.1	LLM	0.99
Bus IV.1	LLM	1.00	Box IV.1	LLM	1.00
Sled I.1	GDINA	0.41	Cart I.1	DINO	0.70
Sled II.1	LLM	0.87	Cart II.1	LLM	0.68
Sled III.1	LLM	0.83	Cart III.1	LLM	0.64
Sled IV.1	LLM	0.74	Cart IV.1	RRUM	0.91

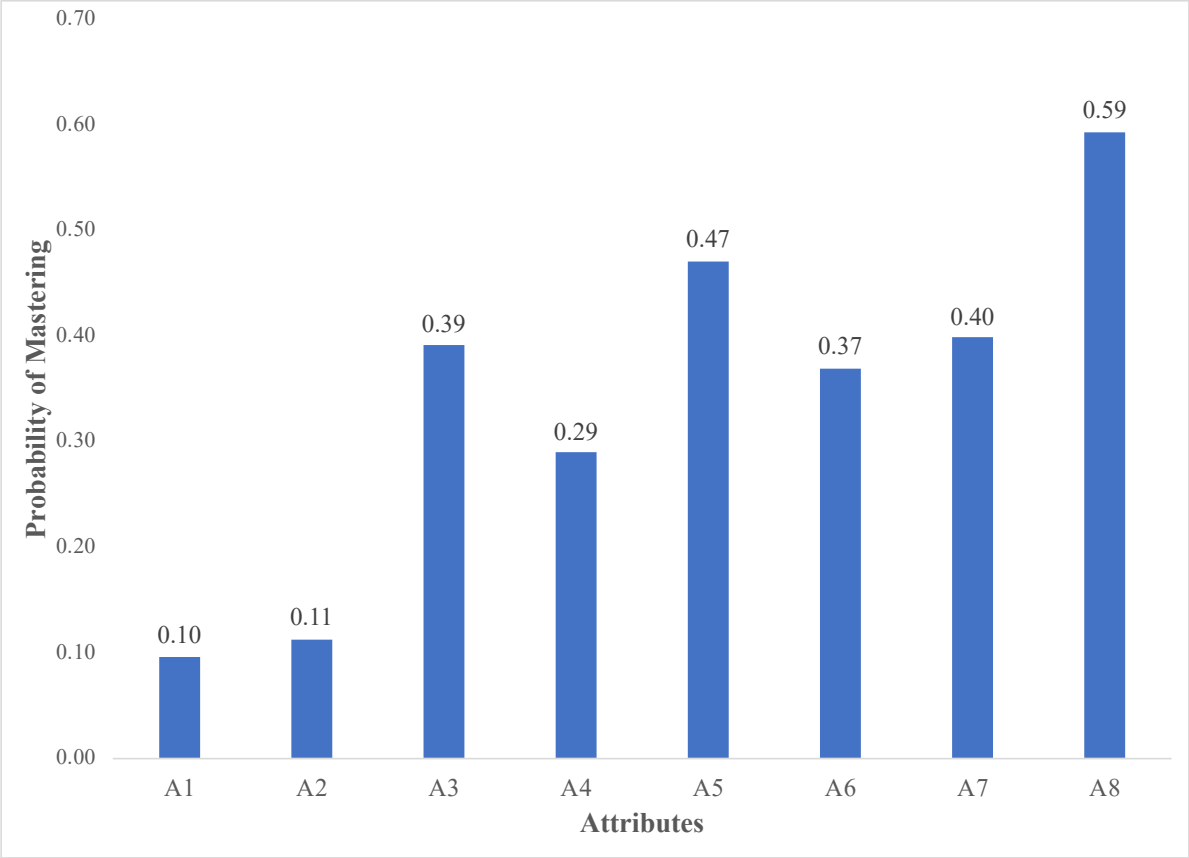


Figure 3.5 Attribute mastery probability for the whole student sample

Item Parameter Estimates. Table 3.6 shows the guessing and slip parameter estimated from the final CDMs. According to Cho (2016), an item with a guessing or a slip parameter higher than 0.6 is considered not having the ability to differentiate students' attribute master levels. All 16 items maintained a relatively low guessing and slip statistics. Students were less likely to guess most of the items correctly without knowing required attributes. The exception is Bus III.1, Sled IV.1 and Cart III.1, which reported a probability higher than the probability of random guessing (25%). It should be noted that all three items tap the understanding about change of motion (either acceleration or deceleration) regardless of what context the item presented. On the other hand, there is a relatively high variation in the slip parameters across all items. Six items reported higher than 25% of probability of missing an item with all required

attributes present, of which Sled I.1 showed the highest slip parameter close to 0.6. The high probability of slip for Sled I.1 could be attributed to a variety of reasons: flaws in item design, use of unfamiliar context (the dogsled context is the least familiar for students), or misspecification of Q-matrix. By carefully reviewing the content of the item and comparing students' response pattern on the item with students' responses to other parallel items (i.e., Bus I.1, Box I.1, and Cart I.1), no obvious flaw was detected in item design which may increase the probability of students slipping and answering the item incorrectly. Further analysis is needed to explore other potential reasons, which includes increasing the sample size to check whether the pattern of high slip probability retains with a different student sample and using cognitive interviews to understand whether the use of unfamiliar contexts triggers alternative thinking which causes students to slip.

Personal Parameter Estimates and Latent Classes. In addition to item parameters, the G-DINA model also estimates individual attribute profile for each student. With respect to individual attribute profiles, students were grouped into 256 latent classes based on the eight attributes. Table 3.7 presents the top 10 latent classes in a descending order based on their posterior probabilities. As the result indicated, the latent class "00001001" had the highest posterior probability (0.121), which means about 12% of the students mastered A5 and A8 while lacking all other attributes. This indicate this group of students have a relatively good understanding of the motion of speeding up and how forces change when the motion changes. The second largest latent class is "00010010" (0.083). This suggested about 8% of examinees were expected to have mastered A4 and A7, which are about constant velocity and friction. The latent classes allow teachers to distribute attention to groups of students rather than individual students, which provides an effective strategy to manage tailored learning in a large classroom.

Besides, the diagnostic feedback for each latent class enables teachers to prioritize their pedagogical focus and refine instructional practices based on students' strength and weakness.

Table 3.6
Guessing and slip parameters for all items

Item ID	**Guessing *(SE)***	**Slip *(SE)***
Bus I.1	0.11 *(0.03)*	0.29 *(0.05)*
Bus II.1	0.00 *(0.00)*	0.00 *(0.00)*
Bus III.1	0.28 *(0.06)*	0.31 *(0.07)*
Bus IV.1	0.00 *(0.02)*	0.00 *(0.00)*
Sled I.1	0.04 *(0.02)*	0.56 *(0.28)*
Sled II.1	0.00 *(0.01)*	0.12 *(0.24)*
Sled III.1	0.09 *(0.04)*	0.08 *(0.04)*
Sled IV.1	0.26 *(0.04)*	0.00 *(0.00)*
Box I.1	0.05 *(0.02)*	0.39 *(0.05)*
Box II.1	0.00 *(0.00)*	0.00 *(0.00)*
Box III.1	0.01 *(0.05)*	0.00 *(0.00)*
Box IV.1	0.00 *(0.03)*	0.00 *(0.01)*
Cart I.1	0.00 *(0.02)*	0.30 *(0.04)*
Cart II.1	0.07 *(0.06)*	0.25 *(0.18)*
Cart III.1	0.32 *(0.07)*	0.05 *(0.02)*
Cart IV.1	0.00 *(0.00)*	0.09 *(0.20)*

Table 3.7
Top 10 latent classes and posterior probabilities

Latent Class	**Posterior Probability**	**Latent Class**	**Posterior Probability**
00001001	0.121	00101101	0.044
00010010	0.083	00000101	0.044
00000001	0.053	00000110	0.034
00100001	0.052	00001100	0.031
00101001	0.050	00010101	0.027

3.4.2 Regressions with Person Parameters as Dependent Variables

Taking the estimated individual attribute profiles as dependent variables (eight attributes per individual attribute profile), we run eight multiple logistic regressions to predict the effect of context familiarity on the probability of mastering a certain attribute with a sample size of 1,478 students. For each regression model, the dependent variable is one of the eight attributes, and the independent variables are individual students' responses to eight survey questions.

Before proceeding to the regression analysis, we first present the result of the simulation study which calculated the classification rates for all attributes overall and for each individual attribute. Table 3.8 shows the classification rate for each attribute. According to the results, overall only 51% of the attributes were classified correctly. At the individual attribute level, all individual attributes reported a similar classification rate which was roughly equal to 0.50. The result of the simulation study indicated an unstable estimation of person parameters. However, as the results of the regression analysis reveal some interesting relations between the effect of context familiarity and the probability of mastering certain attributes, we provide potential explanations for some of the relations (if the relations are true) which may benefit the future studies. Table 3.9 presents the results of eight regression models.

Table 3.8
Classification rate for each attribute

A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
0.51	0.49	0.50	0.52	0.50	0.50	0.48	0.50

As shown in Table 3.9, overall having daily life experience or classroom experience with any of the four cluster contexts did not seem to have a significant effect on the mastery of most attributes. There are a few exceptions. For example, having classroom experience with box (box

survey 2) significantly improved the probability of mastering A2, A4 and A5 (b = 0.64, 0.42, 0.41 for A2, A4, A5, respectively, $ps < 0.01$) with the trend of positive effect on the other four attributes as well. This finding is not surprising. A box (or objects similar to a box, such as a block) is an easily obtained instructional material in physics classroom. Furthermore, in physics a teaching method often involves drawing free body diagrams. Free body diagrams are a method in which the object is drawn (sometimes simplified to a box) and the forces acting on the object are drawn as arrows from the object's center of mass. Therefore, though a box itself may not have been used, the representations the students would have been used to seeing would look like a box. In addition, students may be familiar with physics test items with boxes as many assessment items assessing motion and net force are situated in the context where boxes are objects of interest (e.g., Meltzer & Manivannan, 1996).

Furthermore, the probability of mastering A4 could also be increased if a student had discussed and/or used an example of a dogsled in the physics class (sled survey 2; $b = 0.70$, $p < 0.01$). This finding is somehow unexpected as we suspect that the dogsled is less widely used when teachers explore the relationship between constant velocity and net force. The familiarity indicators in Table 3.3 confirmed that dogsled is the least familiar context to students. We speculate the effect of familiarity with dogsled may be confounded with the context presented. To be more specific, we explicitly mentioned how the brake worked in the sled context, and therefore the frictional force was foregrounded which could have helped students better understand the forces acting on the sled. The students would not have had the interference of a familiar experience to diminish the effect of this foregrounding. Another possible explanation is related to a commonly held misconception by students. One strong misconception about constant velocity is that constant motion requires constant force ("Common ideas about forces and

motion", n.d.). The dogsled has the setting of snowy and icy road which students may be familiar with the idea that the icy surface doesn't involve much friction and an object can move constantly even without any force pushing or pulling. Therefore, it makes sense for students to select the option that an object can move when a net force is zero even though they may fail to apply the concept of the net force. In contrast, the ground conditions in other clusters make it apparent that students cannot ignore the friction. Therefore, students are more likely stick with their misconception.

Unlike the positive effect of familiarity of box and dogsled, having classroom experience with bus (bus survey 2) significantly decreases the probability of mastering A7 (b = -0.50, p < 0.05), which deals with friction. One possible explanation of such an effect may be that though the students are familiar with riding a school bus, they are not familiar with driving a bus/car and therefore are less likely to be familiar with the forces acting on the school bus. In other contexts, such as pushing boxes, students would have had experience with either physically manipulating the objects or, in the case of the sled context, the frictional forces were made more explicit. In the bus context, students were told about the brake and the gas pedal. To them, for the bus to move, the gas must be pushed and for it to slow down the brake must be pushed. Students do not really think about friction because when riding a school bus, a driver is usually maintaining a steady speed, accelerating, or actively braking. Coasting, in which there would be no forward force but a frictional force, is rarely an experience with which a student would have had experience. Furthermore, the act of driving of the bus is largely absent from the students experience as the driver is far from them. Therefore, they only feel the bus move and stop and are unable to connect this motion to the forces involved. We also speculate that as the mass of

the bus is substantial, ideas about momentum may conflate with ideas of force, therefore they neglect friction.

In summary, classroom experience may have a unique effect on the mastery of certain attributes, while daily life experience did not seem to account for any significant variations in predicting students' attribute profiles. As the simulation study indicated an unstable estimation of parameters, further investigations are required to confirm whether the potential effects observed are reliable.

Table 3.9

Results of regressions using students' survey responses to predict probability of mastering A1 to A8

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
Intercept	**-3.07*** **	**-2.74*** **	-0.47	**-1.33*** **	-0.24	**-1.21** **	**-0.85***	0.66
	(0.58)	*(0.58)*	*(0.29)*	*(0.31)*	*(0.28)*	*(0.32)*	*(0.30)*	*(0.29)*
bus survey 1 (daily life)	0.12	-0.70	0.26	0.17	0.03	-0.03	0.11	-0.01
	(0.33)	*(0.26)*	*(0.18)*	*(0.20)*	*(0.18)*	*(0.19)*	*(0.19)*	*(0.18)*
bus survey 2 (classroom)	-0.02	-0.55	-0.43	-0.28	-0.09	-0.10	**-0.50***	-0.04
	(0.28)	*(0.30)*	*(0.16)*	*(0.17)*	*(0.16)*	*(0.17)*	*(0.17)*	*(0.16)*
sled survey 1 (daily life)	0.17	-0.32	0.02	-0.37	-0.14	-0.12	-0.15	-0.07
	(0.35)	*(0.39)*	*(0.21)*	*(0.24)*	*(0.21)*	*(0.23)*	*(0.22)*	*(0.21)*
sled survey 2 (classroom)	0.17	0.42	0.03	**0.70** **	-0.42	-0.11	0.12	-0.41
	(0.34)	*(0.31)*	*(0.21)*	*(0.21)*	*(0.21)*	*(0.23)*	*(0.21)*	*(0.21)*
box survey 1 (daily life)	0.05	-0.18	0.18	0.12	0.03	-0.06	0.21	-0.13
	(0.31)	*(0.28)*	*(0.17)*	*(0.18)*	*(0.16)*	*(0.18)*	*(0.17)*	*(0.17)*
box survey 2 (classroom)	0.23	**0.64** **	0.29	**0.42** **	**0.41** **	0.11	0.30	-0.02
	(0.20)	*(0.20)*	*(0.11)*	*(0.12)*	*(0.11)*	*(0.12)*	*(0.11)*	*(0.11)*
cart survey 1 (daily life)	0.27	0.88	-0.40	0.04	-0.17	0.45	-0.11	0.09
	(0.48)	*(0.53)*	*(0.23)*	*(0.25)*	*(0.23)*	*(0.27)*	*(0.24)*	*(0.24)*
cart survey 2 (classroom)	0.21	0.17	0.02	0.16	0.12	0.06	0.16	-0.10
	(0.20)	*(0.19)*	*(0.11)*	*(0.12)*	*(0.11)*	*(0.12)*	*(0.11)*	*(0.11)*

Note. N = 1,478 students. A1 to A8 were dummy coded with 1 suggesting the mastery of the attribute and 0 otherwise. All survey questions were dichotomized with 1 indicating having related experiences with the object(s) of interest presented in the context and 0 otherwise. The standard errors of estimated regression coefficients were shown in parentheses and italicized. Significant effects were highlighted in bold. The p-values was adjusted based on Bonferroni method. * p <0.05, ** p < 0.01, ***p < 0.001.

3.4.3 Regressions with Item Parameters as Dependent Variables

Multiple linear regressions were conducted with two aggregated familiarity indicators—daily-life experience and classroom experience—to predict guessing and slip parameters, respectively, with a sample size of 16 items. Regression results in Table 3.10 indicate neither daily life experience nor classroom experience were unique predictors of the guessing parameter (b = 0.004, p = 0.958 for familiarity indicator 1; b = -0.04, p = 0.538 for familiarity indicator 2). However, Cohen's f^2 detected a small effect size of classroom experience on decreasing the probability of guessing (f^2 = 0.03). More specifically, if a student has discussed and/or used the object(s) of interest described in the context in the classroom, he/she is less likely to answer the item correctly without knowing any required attributes.

With regards to the slip parameter, having daily life experience or classroom experience with the object(s) of interest in the context reduces the probability of students who master all required attributes respond to an item inaccurately. However, the effect does not show a statistical significance and the magnitude of the effect size is negligible (b = -0.04, p = 0.774 for familiarity indicator 1; b = -0.02, p = 0.894 for familiarity indicator 2).

To ensure the robustness of the results, a simulation study with 30 replications was conducted with the aim to assess the parameter recovery of regression coefficients for the guessing and slip parameters. Table 3.11 presents the statistical summary of the three evaluation criteria. The regression coefficients of the two indicators related to classroom experience recovered well, indicating a reliable estimation of the effects of classroom experience on guessing and slip. With regards to the effect of daily life experience, the results of relative bias suggested an unsuccessful parameter recovery for the regression coefficients of daily life

experience on guessing and slip, especially for the estimated effect of daily life experience on slip.

Table 3.10
Results of linear regressions using two aggregated familiarity indicators to predict guessing parameter and slip parameter

(a) Model estimates for the guessing parameter

	Estimate	**SE**	**adj. *p*-value**	**Cohen's f^2**
Intercept	0.10	0.06	0.240	/
Daily Life Experience (familiarity indicator 1)	0.004	0.08	1.916	0.00
Classroom Experience (familiarity indicator 2)	-0.04	0.07	1.076	0.03

(b) Model estimates for the slip parameter

	Estimate	**SE**	**adj. *p*-value**	**Cohen's f^2**
Intercept	0.19	0.09	0.125	/
Daily Life Experience (familiarity indicator 1)	-0.04	0.13	1.548	0.01
Classroom Experience (familiarity indicator 2)	-0.02	0.11	1.787	0.00

Note. $N = 16$ items. The two aggregated familiarity indicators were dichotomized with 1 indicating having related experiences with the object(s) of interest presented in the context and 0 otherwise. The *p*-values were adjusted based on Bonferroni method.

Table 3.11
Statistical summary of three evaluation criteria for parameter recovery in the simulation study

	Guessing		**Slip**	
	Daily Life Experience (familiarity indicator 1)	**Classroom Experience** (familiarity indicator 2)	**Daily Life Experience** (familiarity indicator 1)	**Classroom Experience** (familiarity indicator 2)
Bias	0.006	0.001	0.132	0.001
Relative bias	1.436	-0.023	-8.499	-0.071
MSE	0.004	0.002	0.021	0.002

In summary, the results of the simulation study provide some evidence that having classroom experience with the context presented in an assessment item may slightly reduce the probability of guessing. However, such context familiarity gained from classroom experience did not impact the probability of slip significantly. With regard to the other source of familiarity—daily life experience, we lack evidence to make any conclusions about the effects of daily life experience on the two item parameters given the unsatisfactory parameter recovery in the simulation study. The results of this part need to be interpreted with caution unless more confirming evidence are collected.

3.5 DISCUSSION

Familiarity with a context is subjective to individuals. Administering a contextualized assessment without considering individual differences in prior experiences could lead to an imprecise measure of students' performance and misleading implications for item development. The present study utilizes a survey to collect inputs directly from students regarding their familiarity with an item context. It is admitted that students' response to the survey questions may be affected by other factors such as motivation, their perceived performance level on the assessment, and accuracy of their memory of what they have learned from the classroom. Still arguably, student self-reports of whether they are familiar with a context, if nothing else, provide the closest measure of the connection between the context and the students.

The primary focus of this study is to explore the influence of context familiarity on two assessment outcomes. As students could gain familiarity with a context from various venues (e.g., culture, religion, habits, and education, etc.), we focus on two general sources—daily life experience and classroom experience. The two guiding research questions examine how and

which of the two sources have an effect on (1) students' mastery pattern of required attributes, and (2) two item parameters estimated from CDM—guessing and slip.

Regarding the first research question, this study detects a potential association between context familiarity and the probability of mastering certain attributes, which requires further investigation. The positive effect of the box context on students' attribute mastery confirmed the argument that familiar contexts may enhance performance (e.g., Song & Bruning, 2016; Washburne & Osborne, 1926a, 1926b). However, the negative association between familiarity of the bus context and mastery of attribute A7 is also in line with the findings of prior work, which claim embedding items in familiar contexts may be detrimental as it may drive students' attention to superficial features of the context instead of the underlying constructs and rely on daily-life experience to make judgement rather than applying construct-relevant skills and knowledge they learned to solve problems (Ahmed & Pollitt, 2007; Salgado, 2017). With regard to the comparison of two sources of familiarity, classroom experience seems to play a critical role in shaping students' individual attribute profile, while daily life experience does not show any significant contribution. Based on this finding, we propose an interesting hypothesis: merely exposing students to a real-life context is not a sufficient condition for learning to happen. Instead, the cognitive engagement with the context probably needs to be coupled with more deliberate instructional effort (e.g., discussed or conducting experiments in physics classes) in order for students to make sense of the underlying concepts and integrate them into their existing knowledge system in a meaningful and coherent manner.

To address the second research question, neither daily life experience nor classroom experience showed significant impacts on the probability of guessing or slip. This finding is not unexpected. Here we provide two possible explanations. First, the number of items is too small

that the present study lacks the power for detecting statistical significance. We conducted a power analysis by calculating the proportion of times the stimulated confidence intervals of regression coefficients do not contain zero during the simulation study. The result yielded a statistical power lower than 0.10 for all four regression coefficients. The second reason for the insignificant impact of context familiarity may be due to the lack of variation in the dependent variables (guessing and slipping parameters) and the two aggregated familiarity indicators. To test the claim, we calculated the Pearson correlation between guessing and slip and the two familiarity indicators. All correlation coefficients were close to or lower than 0.20. Given the limited power, we claim effect size is a more effective measure than *p*-value. Based on Cohen's f^2 index, classroom experience shows a small effect on reducing the probability of answering an item correctly without knowing any required attributes (guessing). Besides, no meaningful relationship was detected between classroom experience and slip (the probability of getting an item wrong with all required attributes present). This finding provides evidence for the advantage of familiar context on test validity. When a student answers an item correctly and he/she is familiar with the context because he/she has encountered similar contexts in physics classes, we are certain that his/her success in the item was less likely due to some factors unrelated to assessed attributes, such as random guessing. This directly addresses the validity of the test score inference by allowing teachers and researchers to interpret test results with more confidence that the test scores reflect students' true ability.

Combining the findings at the person level and item level, this study brings two implications to the research community. First, increasing the exposure of students to various kinds of contexts in classrooms may be beneficial for their learning, as classroom experience is an important channel for gaining familiarity with different contexts and integrating what they

have encountered to what they have learned. Second, the effects of aligning contexts of assessment items with contexts of tasks that are regularly used in classroom practices appear more complex and nuanced, thus item construction should be guided and informed by pilot studies to ensure the contexts do not lead to any construct-irrelevant variance. Contexts that seem resemble what students encounter in their science classes may impact students either positively or negatively, depending on what physics concept or phenomenon is assessed in the item and how it relates to the context.

Besides the content focus on context familiarity, this study is designed with a methodological focus on the application of CDM. Based on the G-DINA framework, this study illustrates the standard procedures to apply a CDM, including defining Q-matrix, model fit, model selection and result interpretation. In addition to the application of CDM, this study takes one step further to explore the relationship between the effect of context familiarity and the two item parameters estimated from CDM, providing deeper insights into how context familiarity impacts item statistics. As the results of model selection revealed potential issues of model misfit or Q-matrix misspecification (three absolute model fit statistics reported conflicting results), which may further impact the reliability of the regression results, we demonstrate the use of simulation studies to evaluate and mitigate the potential risk of unreliable estimation of model parameters. To assess the recovery of regression coefficients on the two item parameters, we performed a simulation study with 30 replications. The results of three evaluation criteria (bias, relative bias, and MSE) revealed a successful recovery of parameters related to classroom experience, while the regression coefficients of daily life experience did not recover well for the two models predicting guessing and slip. Therefore, we acknowledge that the results need to be interpreted carefully and require further investigation.

In summary, this study presents an exploratory approach to explore the role of context familiarity on personal-level and item-level assessment outcomes using the CDM framework. Future work will include continuous efforts on two aspects. The first aspect relates to the conceptualization of context familiarity. First, the current study suggests that students' familiarity with the object(s) of interest may not sufficiently account for their familiarity with the broader item context. For example, students may be familiar with school bus but most of them do not have experience driving a school bus or a car. For future work, we argue the need for conceptualizing familiarity in a more comprehensive way by attending to multiple elements of the context, such as the events described in the context and the role students have interacted with the object. Moreover, the current survey design simplified students' familiarity with a context into a yes-or-no question instead of treating it as a continuum. This issue will be tackled by revising survey questions to be more reflective in capturing students' prior experiences with the objects and events described in item contexts, such as degree of familiarity or frequency students have encountered the object(s) of interest. Furthermore, we plan to incorporate the analysis of student interviews to bring the qualitative evidence in order to unpack how familiar or unfamiliar contexts are perceived and approached by students in their cognitive processes of responding to the items.

The second aspect pertains to methodological efforts to provide a more precise and rigorous measure of the effect of context familiarity. For example, we will extend this study with a large sample size (both participants and items). A large sample will allow us to apply advanced statistical models such as multinomial logistic regression rather than multiple regressions to detect the impact of item context familiarity on person parameter estimates. Second, instead of running a multiple-step analysis (e.g., CDMs and regressions), we will focus on incorporating

the item characteristic variables in the item response function directly to see how they predict the probability of answering an item correctly. What's more, the present study ignores the conditional dependence of items nested in a cluster. Overlooking the local dependence of items may result in aggression bias (i.e., group-level inferences incorrectly assumed to apply to all group members), especially bias in standard error and confidence interval estimates (Fox & Gals, 2016; Sulis & Toland, 2017). Therefore, an important step of future investigation will focus on exploring appropriate psychometric models (e.g., testlet models) to account for the conditional dependence among items. Last but not least, as retrofitting an assessment with a CDM may pose risk of model misfit and Q-matrix misspecification, this study (and the larger research community) will benefit from revising this instrument to improve Q-matrix specification or developing an assessment that is designed for diagnostic purpose. Unlike the current Q-matrix design that requires several attributes to be estimated together (e.g., items assessing A2 tend to also require A6), we plan to develop items that only assess each of the attributes instead of a cluster of attributes.

Chapter 4

Analyzing Sources of Difficulty in NAEP Science Assessment Items: A Machine Learning Approach

Farah Nadeem[1], Dongsheng Dong[2], Min Li[2], & Mari Ostendorf[1]

[1]Department of Electrical and Computer Engineering, University of Washington

[2]College of Education, University of Washington

Abstract

The difficulty level of assessment items is determined by a number of factors, including cognitive demands, item format and linguistic complexity, which have been explored in several studies. This study provides new results and insights into prior work by conducting experiments with a variety of models and model selection criteria in studying the utility of different factors (cognitive demands, item format and linguistic complexity) for predicting item difficulty as measured by aggregate student success rate (*p*-value), leveraging machine learning techniques. Experiments on 132 NAEP science assessment items compares several model selection methods in working with small sample sizes, which is a constraint of publicly available data. Taking a machine learning perspective, we show the benefits of using cross-validation for both model selection and feature interpretation.

4.1 INTRODUCTION

Understanding what makes an item difficult is critical for assessment development. This is particularly true for contextualized items as the inclusion of a context inevitably leads to text and/or non-textual information that test takers need to interact with in their problem-solving process. As more and more science assessment items are situated in real-life scenarios, it becomes more challenging to accurately measure students' understanding as the effect of the underlying constructs often tangles with that related to characteristics of an item, including its context (e.g., linguistic demands and item format). Being able to predict item difficulty and pinpoint potential sources of difficulty in assessment items provides several benefits. From the test validation perspective, it can reveal the potential sources of construct-irrelevant variance and thereby evaluate the extent to which items provide precise and accurate interpretation of students' understanding of the assessed constructs. For item writers, it provides valuable guidance for item development which allows item writers to produce high-quality items in a more efficient way, which is also considered as evidence for the formative stage to evaluate the validity claims for score interpretations (Kane, 2006). From the pedagogical perspective, it offers guidance for classroom assessments in the sense that teachers can select items aligned to specific learning goals and developmentally appropriate for their students so that the assessment results are more readily relevant for teachers to interpret and take action.

Item difficulty has been indexed in multiple ways. In this work, we consider the task of predicting item difficulty in terms of the percentage of correct response to the item (*p*-value), aggregating over a population of test takers, given the text (and possibly visual information) associated with the item. While we know that the probability that a student answers an item correctly depends on both the item characteristics and student understanding level, the data

available in many cases is just the fraction of students who complete the item correctly, so individual student parameters typically cannot be accounted for. In this scenario, as well as when individual student data is available, previous studies have explored a large number of factors that may influence item difficulty (Enright & Sheehan, 2002; Sheehan, Kostin, & Persky, 2006). However, it is challenging to interpret results across studies when model evaluation criteria have such a large range (e.g., the percentage of variance explained, R^2, varying from 0.10 to more than 0.90), making it difficult to identify specific findings that generalize across multiple studies. This is in part due to the fact that studies look at different sets of item characteristics, different types of items, and different populations of test takers. In addition, we argue here that the methodologies often used are limiting the generalizability of findings. The goals of this work are to provide insights into the findings from previous work by shedding light on issues associated with data analysis methodology, as well as to provide new results related to specific item characteristics.

This paper uses a machine learning framework to look at the combination of item format (response type), cognitive demands (topics and practices) and linguistic complexity features drawn upon characteristics that have been explored in various forms in several prior studies. Broadly, our approach is to analyze student performance on science assessment items with multiple models, different model selection criteria and alternative criteria for interpreting importance of features. We compare methods that have been used in prior work on item difficulty prediction with related approaches that are standard in machine learning. In particular, we leverage the practice of holding out data in assessing both model fit and feature importance.

This work makes two main contributions. First, we outline a methodology for identifying item characteristics that are associated with aggregate student performance that provides results

that better generalize across data sets and lead to more consistent findings about feature importance. Second, we provide results on item difficulty prediction that confirm and clarify prior findings about item characteristics, specifically item format, cognitive demands and linguistic demands.

The remainder of the paper proceeds as follows. In Section Background, we provide context for this study with a discussion of prior work in the education literature and foundational methods from machine learning. Section Methods outlines the methods used for modeling, model selection and interpretation. Results and analyses are presented in Section Results. Finally, Section Discussion summarizes the contributions and limitations of the study.

4.2 BACKGROUND

In this section we look at existing research on item difficulty prediction that informs our study, pointing out commonalities and differences in findings associated with these studies. We then cast the problem of modeling the effect of item characteristics on student performance in a machine learning framework and provide links between terminology in the different research communities. In this context, we explore how machine learning techniques can be used to strengthen the methodology for item difficulty prediction.

4.2.1 Prior Research on Item Difficulty Analysis

Previous studies have explored a large number of factors that may influence item difficulty including cognitive demands, item formats, item topics, linguistic demands and so on. These features are referred to as item response demands by Ferrara and Duncan (2011), which encompass the content, cognitive and linguistic knowledge and skills required to solve or partially solve an item. Based on the item response demand framework, Ferrara, Steedle and

Frantz (2018) reviewed 24 studies on item difficulty modeling and summarized item response demands identified as potential predictors of item difficulty. According to Ferrara et al. (2018), 15 out of 24 studies employed linear regression as the major methodological approach and yielded a varying range of R^2 values.

As the review by Ferrara et al. (2018) includes only two studies related to science assessments, we composed a more STEM-focused literature review by keeping the science and mathematical studies reviewed by Ferrara et al. (2018) and adding more research on science assessments. We summarized 10 studies on item difficulty modeling, separating studies that do and do not incorporate student characteristics in Tables 4.1 and 4.2, respectively. As noted in the tables, the studies use different parameters for item difficulty. In our study we use *p*-value, since it is available for more NAEP items and it aligns with the dependent variables used in prior work on NAEP assessment items (Valencia, Wixson, Ackerman, & Sanders, 2017). Most of the studies use linear regression to predict item difficulty, not noted in the table for brevity. An exception is Crisp and Grayson (2013), which uses a linear logistic test model (LLTM) as well as linear regression. In addition, most studies report R^2 or adjusted R^2 as an indicator of model fit, using all available data to fit the model. The number of items used in the studies ranges from 18 to 216.

The tables also note the predicted variable and the student/item characteristics explored as predictors, highlighting those characteristics found to be significant for predicting item difficulty. Since these studies test different sets of item parameters on different item collections, it is impossible to compare specific findings across studies. Further, there is no one characteristic explored in all studies. However, it is informative to look at findings for broad categories of

features. Here we present examples related to the different item characteristics explored in our work. We observe that there are some conflicting results in these findings.

Most studies look at some type of cognitive demands and find that at least one factor is significant. The work by Crisp and Grayson (2013) demonstrates two common approaches to model the effects of cognitive demands. With 40 items from a UK-developed international A-level physics examination, a total of 27 features of the assessment items were identified and grouped into four distinct categories: physics knowledge and understanding, question processes like recalling equation or unit and recalling physics concepts, cognitive complexity of the item such as the number of components, operations or ideas and the links between them, and observable attributes including the total amount of reading, maximum sentence length, and density of technical physics words. As the first step of the analysis, they fitted a linear regression using the 27 features as predictors. The predicted variable was the difficulty estimates obtained by fitting the response data to a Rasch simple logistic model. The regression analysis reported a R^2 value of 0.89 and an adjusted R^2 value of 0.66. Four of the item characteristics were identified to have significant effects on Rasch item difficulty estimates: total amount of reading (question attributes), the requirement to employ physics concepts (question process), work with symbols (question process), and carry out calculations (question process).

Several studies look at item format (El Masri, Ferrara, Foltz, & Baird, 2017; Mesic, 2011; Mesic & Muratovic, 2011; Le Hebel, Montpied, Tiberghien, & Fontanieu, 2017; Wright, Eddy, Wenderoth, Abshire, Blankenbiller, & Brownell, 2016), primarily response type, sometimes referred to as openness, but some also look at the use of graphics. Studies distinguish between various forms of constructed response (CR) and multiple-choice items, where CR might include short answer, fill in the blank, essay, graphing, or drawing. Most studies found response type to

be significant when included with some form of cognitive demand features. For example, the study by Le Hebel and colleagues (2017) examined the effects of four item features on 103 PISA 2006 science items, including the cognitive demand based on Webb's (2007) Depth of Knowledge (DOK) level, the dependence on or independence of the information provided in the unit and/or item introduction, the item format such as open-ended response and multiple choice, and the competency level specified by the Program for International Student Assessment (PISA). Their analysis of variance (ANOVA) models show an R^2 of 0.30 for predicting item difficulty. Only cognitive demand (indicated by DOK level) and item format proved to be statistically significant predictors of items' level of difficulty. Specifically, open-ended responses were the most difficult item format and multiple choice were the least difficult. Another recent example is the work by El Masri and colleagues (2017), who collected responses to 216 science items from the Key Stage 2 national science assessment and examined the effects of five types of item characteristics on item difficulty. The independent variables included curricular variables, item formats, depths of knowledge, nature of stimulus and language variables. The Item Response Theory (IRT) model threshold parameters were estimated and used as the dependent variable in regressions. They found that extended constructed items increased item difficulty, in line with the work of Le Hebel et al. (2017). Overall, the regression models report R^2 of 0.23, with extended constructed items and the presence of photos found to be significantly predictive of item difficulty.

Studies incorporating linguistic complexity[10] have not led to consistent findings. Studies look at some form of linguistic complexity, including measures of reading level (Höttecke, Feser,

[10] We consider linguistic complexity as opposed to linguistic demands, since we focus on difficulty associated with language use, as opposed to higher demands posed by longer texts. As such, we do not consider total word count to be a linguistic complexity feature.

Heine, & Ehmke, 2018; Rosca, 2004) and specific variables hypothesized to distinguish between reading levels (Crisp & Grayson, 2013; El Masri et al., 2017). Only Höttecke et al. (2018) found that this factor was significant for predicting item difficulty, using a simple characterization of linguistic complexity (low, medium and high). Each of the six multiple-choice items used in the study were modified to be of low, medium and high linguistic complexity, and student performance data was collected for the resultant 18 items (three versions for each of the six items). The results showed that the effects of linguistic complexity on item difficulty was not consistent across all items. The one-way ANOVA showed only two of the six items' scores could be significantly predicted by level of linguistic complexity. One item showed a reversal effect: low linguistic complexity tended to increase the level of item difficulty. El Masri et al. (2017) suggested that the lack of efficacy of using linguistic features for assessment items in their own study might be due to their features (Coh-Metrix variables (Graesser, McNamara, & Kulikowich, 2011)) being less useful with short passages, which provide limited information in computing these statistics. This effect was also seen in a study on predicting language difficulty in science assessment items (Nadeem & Ostendorf, 2018).

From our survey of previous work, we see that typically models for item difficulty prediction are trained with very small numbers of samples, and metrics are reported on the data the model is trained on. This leads to the case where we have findings that hold for one data set but may not hold for another set of items. From Table 4.1 this outcome is reflected in that different features are flagged as significant in different studies. We also observe that R^2 and adjusted R^2 have been extensively used as the most important (and sometimes the only) criterion to evaluate models and justify results. Gayawan and Ipinyomi (2009) used simulated data to compare the relative performance of Akaike information criterion (AIC), Schwarz information

criterion and adjusted R^2 using fertility models. Results showed that adjusted R^2 may have the risk of over parametrization compared to the other two criteria as it consistently chose complex models. To counter these issues, in this work we propose approaches from machine learning to strengthen existing methodologies for model training and evaluation, and we run commonly used statistical analyses to put our work in context with existing studies. Using these methods, we further explore linguistic complexity as an indicator of item difficulty.

Table 4.1

Existing studies on aggregated item difficulty prediction, listing number and type of items, predicted variable, and predictors used in the study

Study	**Items**	**Predicted Variable**	**Predictors *(italicized when significant)***
Mesic & Muratovic (2011)	123 physics (66 multiple choice, 57 constructed response)	Rasch item difficulty	*Item openness; interference effects of intuitive and formal physics; relationships; related relationships; experimental method; mitigating factors (e.g., whether the item can be solved by remembering fragments of knowledge); analytic representation*
Mesic (2011)	123 physics (66 multiple choice, 57 constructed response)	Item discrimination power measure	*Item openness; relationships; related relationships; interference effects of intuitive and formal physics; analytic representation; number of depictors; grade level of item; combined features (grade × relationships, grade × related relationships, item openness × relationships, item openness x analytic representation)*
Crisp & Grayson (2013)	38 physics (multiple choice)	Rasch item difficulty	Question attributes (*total amount of reading*, maximum sentence length, concepts, context, visual resources, importance of options), question processes (e.g., recalling equation or unit, *using physics concepts*, selecting equation or data, *working with symbols*, *calculating*), physics knowledge and understanding (e.g., scientific phenomena, scientific applications), and cognitive demand (e.g., complexity, abstractness, response strategy)
Rosca (2004)	104 science (multiple choice)	Rasch item difficulty	Presence of a figure, Flesch reading level, the *mean number of words in distractors*, number of options, *ratio of number of words in correct option and mean number of words in distractors*, and *cognitive level*
El Masri et al., (2017)	216 UK science (objective & short constructed response)	2-parameter graded response model threshold parameters	Curricular variables (e.g., topic, subtopic, and concept), item type (e.g., *extended construct response*), depth of knowledge, nature of stimulus (i.e., text, *photo*, graph, schematics representation), and language variables (five dimensions from Coh-Metrix software)
Le Hebel et al., (2017)	103 PISA science (objective & constructed response)	*p*-values	*Depth of knowledge,* necessity of context information, *item format,* and PISA competency

Turner et al. (2013)	48 PISA math	Item difficulty	*Communication*; *devising strategies*; mathematizing; representation; *using symbolic, formal, and technical language and operations*; and *reasoning and argumentation*
Morrison & Embretson (2014)	Math (number unspecified)	Item difficulty	19 attributes in 5 cognitive competencies: *translation (e.g., modifier prop); integration (e.g., translating word equation); solution planning (e.g., number of subgoals); solution execution (e.g., number knowledge);* and *decision processing (e.g., decision processing confirmation)*

Table 4.2
Existing studies on predicting individual student performance

Study	**Items**	**Predicted Variable**	**Predictors *(italicized when significant)***
Wright et al. (2016)	87 biology exams, 4810 students	Student test score	Exam: weighted Bloom's index of items; proportion of constructed-response; *weighted difficulty index* Student: course; *time; cumulative GPA;* gender; SES Combined: *gender × weighted Bloom's index of items; gender × proportion of constructed response; SES × exam characteristic; SES × proportion of constructed response*
Höttecke et al. (2018)	6×3 physics items, 1346 students	Student item score	Item: *level of linguistic demands* Student: grade in Math; *migration background; number of books at home*, German reading proficiency (self-estimated); *grade in German;* grade in Physics; *c-test deviated from the canonical c-test concept*

4.2.2 Machine Learning Methodology

In this section, we outline standard machine learning methods for approaching a prediction problem, and specific models used in this study, relating terminology and methods to that used in prior work on item difficulty analysis.

Training and Evaluation. The problem of predicting item difficulty (represented here in terms of the item *p*-value) given the text of the item can be framed as a machine learning problem, where the task is to learn a function $\hat{p}_i = f(x_i)$ from a collection of examples $\{(x_i, p_i); i = 1, \ldots, N\}$ (the training set), where x_i is a vector of features (explanatory variables[11]) for the *i*-th item, and $p_i \in [0, 1]$ is the item difficulty (dependent variable). Assuming a particular functional form for $f(\cdot)$, the parameters of the function are learned by minimizing a loss function on the data, often referred to as "training." Model training corresponds to model fitting, i.e., finding the parameters that best fit the data, and the loss function corresponds to the negative version of the goodness of fit criterion.

In machine learning, an important issue is overfitting. As the power of the prediction function is increased (either by adding features or using more complex functions), it may be possible to perfectly fit the training data, but then the model will not generalize well to new data. More generally, the goodness of fit tends to be better for the training data than on data that was not used in learning the model parameters. In other words, the goodness of fit on the training data is optimistically biased. Even relatively simple models can overfit when the training set is small. In order to get an unbiased estimate of goodness of fit, it is standard practice to assess the learned predictor on data that it is not trained on, which is referred to as the evaluation or test set.

[11] "Independent variables" is a commonly used term, but these variables are not always statistically independent, and relying on that assumption can sometimes be problematic.

Ideally, if there is a large amount of labeled data available, a held out test set is used. However, when there is not enough data for a dedicated held out test set, cross-validation (CV) is used. In the studies described previously, data set sizes vary from 18-216, a range that benefits from CV. In a typical m-fold CV setting, the data is first split into m subsets (folds). One fold of the data is held out, and the model is trained using each of the remaining $m - 1$ folds of data, then evaluated on the held out fold. This process is repeated m times, each time with a different fold held out. The final evaluation score is the average of scores on the held-out folds. This allows the use of all the data in both training and evaluating the model, at the cost of having to learn multiple models.

The processes of model selection and feature selection can also result in overfitting, and a variety of strategies have been introduced to counteract this. In general, these strategies involve using a loss function that is related to but somewhat different from the final evaluation criterion. One approach is to add a penalty function to the loss function. Examples commonly used in educational measurement models are AIC and BIC, where a maximum loglikelihood objective is combined with a penalty term that accounts for model complexity and training set size. Specifically, the AIC and BIC are defined as $2k - 2\log(\hat{L})$ and $\log(n)k - 2\log(\hat{L})$, respectively, where $\hat{L}$ is the likelihood of the data given the trained model, n is the number of samples, and k is the number of model parameters. Adjusted R^2 can be motivated by the same idea. Another approach to avoid over-fitting is to use a parameter regularization penalty, where the penalty weight is chosen using held-out data. In a limited data scenario, holding out data to learn the penalty weight further reduces the training data in the CV set-up. While use of penalty terms in model fitting and model selection reduce the problem of overfitting, they do not eliminate the problem of bias in the performance estimate. In addition, the different penalty

terms reflect different statistical assumptions about the data, which often do not hold in practice. For these reasons, it is important to use these criteria in addition to – and not in place of – evaluating performance on independent data. A difficulty in interpreting prior work in modeling item difficulty is that most studies do not report results on held-out data.

In this paper, we report results using a variety of model selection criteria, and demonstrate the importance of using held-out data in item difficulty analysis, where publicly available data sets are small.

Prediction Models. The problem of predicting item difficulty can be posed in different ways. If it is posed as predicting a continuous variable $p_i \in R$, it is a regression problem, and the evaluation criterion is typically mean-squared error (MSE) or the normalized version, R^2. An equivalent solution can be obtained by using a negative log likelihood loss function with the assumption that $p_i = f(x_i) + v_i$, where v_i has a Gaussian distribution. If the problem is posed as modeling the probability of a binary outcome $p_i \in [0, 1]$ (whether or not the item will be answered correctly), a standard evaluation criterion would be cross- entropy or normalized cross-entropy. In either case, defining a model requires specifying the form of the prediction function $f(\cdot)$.

For most of the results presented here, we model item difficulty using regression. If the function $f(\cdot)$ is linear, then the approach is referred to as linear regression. We also explore decision tree functions, which are defined by a sequence of binary questions about the features.

Since *p*-values naturally represent a binary variable $p_i \in [0, 1]$, we also experimented with a logistic regression model, in which case $f(\cdot)$ is a loglinear function. The loglinear model has the same mathematical form as the LLTM; however, it will be used here for predicting scores aggregating over all students vs. accounting for individual performance.

4.3 METHOD

A general approach to understanding the impact of different item characteristics on item difficulty is to analyze the impact of these characteristics used as features in predicting the percentage of correct response to an item (*p*-values). In this work, we apply this methodology using multiple models, model selection methods and interpretation methods, in order to provide more reliable findings given the small data set available. In this section we present the data, models and model selection methods, and interpretation techniques associated with the different models.

4.3.1 Items

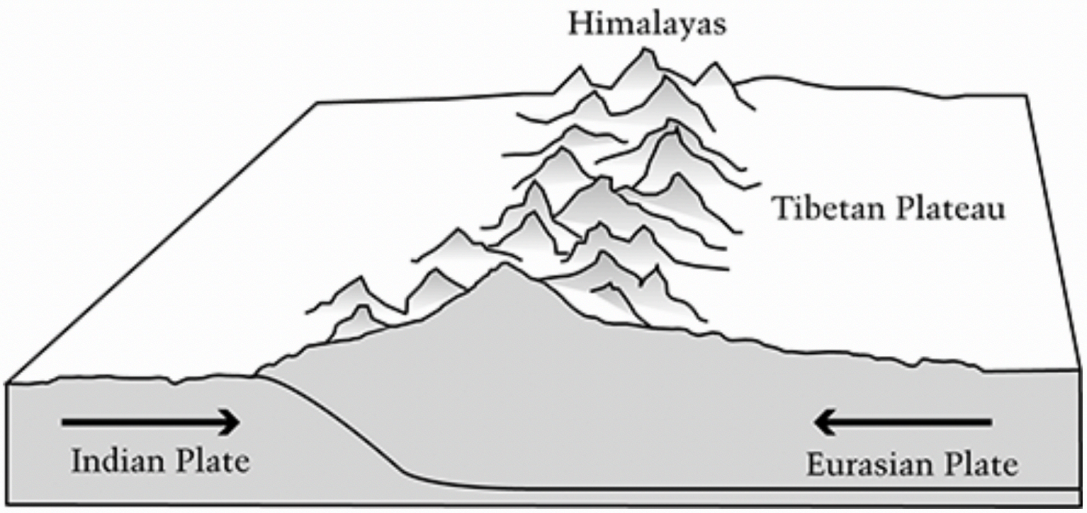
Our data set consists of 8th grade released science assessment items from the National Assessment of Educational Progress (NAEP, https://nces.ed.gov/nationsreportcard). We collected 132 items, for the years 2000-2011, each with the associated *p*-value. For items used in multiple years, the *p*-value of the item is averaged over the years the item has been administered.

The associated meta-data for each item includes the response type, content classification (topic), and science practice. For response type, which corresponds to an item format feature, each item was characterized as either being multiple choice (MC), short constructed (SCR) or extended constructed response (ECR). Each item was associated with one of three content topics (earth and space sciences, life science and physical science) and one of three science practices (identifying science principles, using science principles and using scientific inquiry). Item counts for the different meta-data characteristics are shown in Table 4.3; Figure 4.1 shows a sample item.

Table 4.3
NAEP data—132 items broken down by content, practice and response type

Content classification		Science practice		Response type	
Physical science	39	Using science principles	64	Extended response	17
Life science	52	Identifying science principles	37	Short response	54
Earth & space science	41	Using scientific inquiry	31	Multiple choice	61

1. The diagram below shows the collision of two tectonic plates in Asia.



What is a result of this collision?

A. Volcanoes erupt periodically.

B. The Tibetan Plateau slowly sinks.

C. The Himalayas increase in height each year.

D. Glaciers on the Tibetan Plateau melt.

Figure 4.1 Sample NAEP item. Content topic is earth and space science; practice is using science principles; and response type is multiple choice.

In addition to NAEP, we also used released science assessment items from 2000 to 2015 from PISA. We collected 48 items with associated *p*-values. The associated metadata for each item includes response type (i.e., multiple-choice, short and extended constructed response) and whether the item has a visual element or not. Each item was hand-coded in terms of competency (e.g., explain phenomena scientifically) by an assessment expert specializing in science education who manually mapped the different competency schemes across different years in PISA to the 2015 PISA competency framework. Since we were only able to collect a small number of PISA items with *p*-values, this data is only used for analysis of linguistic features, specifically to explore the impact of passage length, since the PISA items tend to have much longer texts.

4.3.2 Features

The meta-data labels associated with each item were mapped to binary indicator variables for use in the different models for predicting item *p*-values. For example, the topic label was associated with one indicator each for earth and space science, life science, and physical science. The result is 3 binary format features and 6 binary cognitive demand features (3 topics and 3 practices).

For linguistic complexity, we followed a multi-step process to identify a small number of candidate features using data sources other than the NAEP items, since there is a potential for overfitting when selecting features directly on the full item data set. In particular, features chosen based on correlation with NAEP item difficulty are optimistically biased, and they may not generalize to other data. In brief, the approach used here is to assess and select from a large number of features that have been proposed in the literature, evaluating them in terms of utility in predicting grade level of K-12 science texts and correlation with grade-level of non-NAEP

science items. Details of the selection process are provided in Appendix D Selection of Linguistic Features. The selected features include:

- the average age-of-acquisition (AoA) for all words in an item using word scores pulled from a psycho-linguistic database (Cortese & Khanna, 2008);
- the ratio of the log of unique words to the log of all words in an item, transformed using a function learned for predicting grade level of a text;
- the predicted grade level from a pretrained neural network model (NN) from (Nadeem & Ostendorf, 2018); and
- the average word length in characters.

These features are computed with the full text of the items, including answer options when they are present.

Since we have a small set of items to train the models, the initial model selection work uses only the feature that was most correlated with the non-NAEP science assessment items, which was age-of-acquisition (AoA). In subsequent experiments with a subset of models, we compare the four linguistic features.

4.3.3 Models and Model Selection

In order to convincingly demonstrate that specific features impact item difficulty, it is important to show that results hold on data that has not been used to fit the model. To support our claim that cross-validation is more useful than other approaches, we conduct experiments with multiple model selection criteria, described in Section Model Selection. In addition, we explore multiple models, as described in Section Models to look for consistency of results across models.

Model Selection. A popular criterion for evaluating model fit when predicting item difficulty is R^2, which indicates how much variance is explained by the model (higher R^2 is

better). However, when R^2 is reported on the full data set, it is optimistically biased and sensitive to overfitting, so most studies select the best model and report results based on some other criterion.

The criterion that this paper advocates for is average R^2 of the CV test sets using m-fold CV, referred to here as CV-test R^2. Specifically, we split the data into $m = 3$ sets of length 44 items. For each of the 3 models we train, we combine two splits for training and use the third for testing.

For comparison, we also compute three other indices that have frequently used for regression model selection in the education literature: adjusted R^2, AIC and BIC, all of which include a correction term for models with more parameters. AIC and BIC are minimized, so lower scores correspond to better models. For AIC and BIC, we use the standard implementation in StatsModels[12], which assumes that the MSE provides a good estimate of the noise variance needed for computing the log likelihood. This assumption is reasonable if the model form is approximately correct and there are sufficient data samples, but it is potentially problematic for small data sets, as our results will show.

To test whether the difference between two models is significant, we use bootstrapping, based on the CV R^2 scores. For two models being compared, we first obtain the model results for all three CV test folds (total of 132 items) for both sets of trained models. We then randomly sample $n = 132$ items with replacement (bootstrapping) and compare the model performance for these two sets. This process is then repeated 10,000 times, and significance is taken as thefraction of times the model with the lower score on the entire data performs better than the model with the higher score on the entire data.

[12] http://www.statsmodels.org/stable/

Models. We fit three types of models for predicting the item *p*-values: multiple linear regression, log-linear model (binary logistic regression), and decision tree regression (Breiman, Friedman, Olshen, & Stone, 1984). The features $x = [x_1 \ldots x_d]$ are predictor variables, and the target variable for prediction is $p \in [0, 1]$.

Multiple linear regression is the most commonly used model in predicting item difficulty. The *p*-values are predicted using a linear combination of features:

$$\hat{p} = b_0 + \sum_{i=1}^{d} b_i x_i \tag{4.1}$$

The linear regression model parameters $\{b_i\}$ are chosen to minimize the mean-squared prediction error on the training set.

For binary log-linear models, the *p*-value is estimated as a logistic function of a linear combination of features:

$$\hat{p} = \frac{1}{1 + e^{-(b_0 + \sum_{i=1}^{d} b_i x_i)}} \tag{4.2}$$

The log-linear model parameters are chosen to maximize the log likelihood of the training data.

Decision tree regression models involve asking a series of questions about the features, following a tree structure, and then assigning a prediction according to the final leaf of the tree that the question answers lead to. The tree structure is learned in a greedy fashion, by finding the partition of the feature that gives the maximum reduction in mean squared error, and repeating this until either the maximum specified size is reached, or there are no more partitions that improve the MSE.

When given a large number of features, or continuous-valued features (which can be associated with a large number of questions), decision trees are susceptible to overtraining. To counteract this problem, it is common to use some form of pruning or depth selection. In this

work, which is based on the Scikit-learn software package (Pedregosa et al., 2011), we use cross-validation within the training data to select the maximum depth of the tree before training the final tree. Depth selection is important because a large tree depth can lead to overfitting on the training data. The specific implementation of depth selection involves choosing the depth that minimizes 5-fold CV development test[13] mean squared error using the training set, then retraining the model on the full training set with the chosen depth. To study the impact of controlling the tree depth, we train decision trees both with and without depth selection, as well as with and without CV testing. When using depth selection and CV testing, the CV associated with depth selection operates separately on each training configuration in the overall 3-fold CV.

4.3.4 Interpretation

The underlying goal of this work is to determine which item characteristics are important for predicting difficulty of the item. To that end, we explore multiple methods: i) significance tests for coefficients in linear regression analysis, ii) comparison of different feature sets using model selection methods, and iii) decision tree methods.

For linear regression, we compute the standard error for each coefficient in linear regression, which is an estimate of the standard deviation of the coefficient. The *t*-statistic is computed by dividing the coefficient value by the standard error, which is compared to the Student's t distribution with $n - k$ degrees of freedom to give the *p*-value. To support the claim that tests run on the full data set can be biased, we also compare regression models trained with different linguistic features and report different model selection criteria described in section Model Selection. In addition, we implement an alternative strategy for estimating variation of the

[13] The term development test indicates that held out data is used but that it is different from the actual test data.

coefficients, using random sampling of the training set to learn multiple regression models. The resulting distributions of coefficients provide insights into relative difficulty associated with all different item types.

The decision tree regression model provides an alternative method of identifying important features, since the learning algorithm involves a greedy search for the features which have the greatest impact on prediction. Further, the depth selection process eliminates questions that provide insignificant benefit. By doing sequential tests, the decision tree feature selection mechanism accounts for dependence between features to some extent.

Prior studies (and this work) have found that linguistic complexity features are not significant, though one might expect that linguistic complexity should matter. It may be that this factor is sufficiently well controlled for in item design, but we hypothesize that there could also be a confounding issue of interdependence of linguistic and cognitive and/or response type features. To explore this question, we compute the correlation between the linear regression coefficients of format and cognitive features with the average values of the linguistic features for the training items, as described in more detail in Section Results.

Lastly, we conduct experiments to test the hypothesis that linguistic features are less useful when the items are short. We assess the efficacy of linguistic features (computed for the entire item text) as predictors of item difficulty for short and long items by comparing *p*-value prediction for NAEP items, which tend to have shorter contexts, with PISA items, which typically have longer contexts. We report results on model performance for both sets of items with and without using linguistic features to compare how the performance changes in the case of longer items. In these experiments, we restrict the set of NAEP cognitive demand features to roughly match what is available for the PISA items.

4.4 RESULTS

In this section, we present results for model selection, identify features that have a significant effect on prediction performance, and provide analyses to better understand the lack of significance of linguistic features on item difficulty.

4.4.1 Comparison of Model Selection Methods

We first present the results for item difficulty prediction using multiple linear regression, a log-linear model and decision trees (with and without depth selection) for two sets of explanatory variables: 1) item format and cognitive demands, and 2) item format, cognitive demands and linguistic complexity as characterized by average age-of-acquisition (AoA) of words in an item. Models are trained to fit all of the data as well as using 3-fold CV, where the same CV split is used for all conditions.

For each condition, Table 4.4 shows the R^2 on the training data, using all the items to train the model, and the average R^2 on the train and test folds for 3-fold CV. For CV, the results on the training set (CV-train) are the average of the 6 training folds in the three different train/test splits, and the results on the test set (CV-test) are the average of the three test folds across the three splits. As expected, the results for the training data (both All and CV) are better than on the test data. Based on the CV-test results, the decision tree with depth selection is the best model, and the model performance is significantly better than linear regression ($p < 0.05$) using bootstrap. The improvement in R^2 from adding the AoA linguistic feature to the decision tree is not significant. Because the AoA feature is continuous valued, the decision tree that uses it without depth selection is able to perfectly fit the training data, but the overfitting results in a predictor that is worse than a constant predictor (sample mean) on independent test data. The detailed breakdown of the results for the different folds is given for the two decision trees with

all features in Table 4.5. In general, the R^2 on a particular fold is higher when it is used in training than in testing.[14] The R^2 result for the decision tree with depth selection fit to the full training set is closer to the CV test result than the linear and logistic regression models, because the process of depth selection itself involves cross-validation.

Table 4.4
R^2 for different prediction models and feature sets, comparing results when fitting to the full data set vs. 3-fold CV, showing that the training R^2 is not a reliable indicator of test R^2

Features	Model	All-Train	CV-Train	CV-Test
AoA, Item format, Cognitive demands	Linear regression	0.30	0.32	0.17
	Log-linear model	0.30	0.32	0.17
	Decision tree w/ DS	0.38	0.40	**0.32**
	Decision tree w/o DS	**1.00**	**1.00**	-0.16
Item format, Cognitive demands	Linear regression	0.30	0.30	0.20
	Log-linear model	0.30	0.32	0.18
	Decision tree w/ DS	0.38	0.40	**0.30**
	Decision tree w/o DS	**0.49**	**0.55**	0.14

Note. The best result for each configuration is highlighted in bold. Decision tree w/ DS refers to the decision tree with depth selection. Decision tree w/o DS refers to the decision tree without depth selection.

Table 4.5
R^2 for decision tree with and without depth selection using all features

(a) With depth selection

	Fold 1	Fold 2	Fold 3
Train 12	0.48	0.38	*0.27*
Train 23	*0.41*	0.38	0.30
Train 13	0.49	*0.28*	0.34

(b) No depth selection

	Fold 1	Fold 2	Fold 3
Train 12	1.00	1.00	*0.15*
Train 23	*-0.16*	1.00	0.99
Train 13	1.00	*-0.49*	0.99

Note. The performance of the fold used in testing is indicated in italic.

[14] Because of the statistical variation across samples (folds), it is possible for a test fold to have higher R^2 than the folds used in training.

From the above results, it is clear that using R^2 when fitting to the full set of data can give misleading results. For that reason, many studies use statistics that account for the number of free parameters in the model and the number of sample points, including adjusted R^2, the Akaike Information Criterion (AIC) and/or the Bayesian Information Criterion (BIC). For these statistics, higher R^2 and adjusted R^2 are better, and lower AIC and BIC scores are better. Table 4.6 presents these statistics for the different regression models (not including log-linear models), using the implementation in StatsModels (Perktold, Seabold, & Taylor, n.d.). We see that all three criteria still give an overly optimistic assessment of performance for decision trees without depth selection when the continuous-valued age-of-acquisition feature is used, since assumptions behind these criteria breakdown when there is severe overfitting.

In summary, in experiments with multiple models and model selection methods, we find that decision tree with depth selection is the best performing model and that there is no significant benefit from using the AoA linguistic feature. More generally, we show that R^2 reported on the entire data set is particularly susceptible to overfitting and, additionally, that the model selection criteria often used to compensate for this are not as reliable as cross-validation. The assumptions that these criteria rely on simply do not hold for the small data sets used here, particularly for decision trees. Since feature selection can be thought of as a form of model selection, the problem of overfitting on the full data set also impacts methods for identifying significant features, as shown in the next section.

Table 4.6
Metrics for all regression models

Features	Model	R^2	Adj R^2	AIC	BIC	Test R^2
AoA, Item format, Cognitive demands	Linear regression	0.30	0.26	-73.1	-48.30	0.17
	Decision tree w/ DS	0.38	0.36	-95.20	-79.00	**0.32**
	Decision tree w/o DS	**1.00**	**0.99**	**-589.50**	**-366.20**	-0.16
Item format, Cognitive demands	Linear regression	0.30	0.27	-74.70	-52.80	0.20
	Decision tree w/ DS	0.38	0.36	**-95.20**	**-79.00**	**0.30**
	Decision tree w/o DS	**0.49**	**0.42**	-93.80	-37.10	0.14

Note. The best results are highlighted in bold. The best performing model on the entire data does not give the best performance on the test set.

4.4.2 Feature Selection

Linear regression models have been used in many studies aiming to identify characteristics of items that are significant for predicting student performance. In Table 4.7, we report standard feature analyses associated with linear regression, providing the coefficient *B* and the *t*-statistic for models learned on the full data set with three different sets of features. Features identified as significant (based on the assumption of independent input variables) are indicated with asterisks. When no linguistic features are used, we find that the format, topic and science practice features all have significant factors, broadly consistent with prior results in the literature. Adding the age of acquisition linguistic feature reduces the significance of some of the cognitive features. When all linguistic features are included, only the average word length is significant. As will be shown in the next section, the average word length feature has a reasonably high correlation with the linear regression coefficient associated with both science practice and item format features, so the assumption of independence in the significance tests does not hold. This interdependence of features combined with overfitting results in a misleading lack of significance for the more important cognitive and format features.

Table 4.7
Multiple linear regression analysis results for models trained with format and cognitive features only (left), these features plus the AoA linguistic feature (center), and all linguistic features (right)

Variables	***B***	***t***	***B***	***t***	***B***	***t***
Intercept	0.22**	22.36	0.26*	3.31	0.48	0.78
Linguistic Feature						
Age-of-acquisition	-	-	-0.02	-0.06	-0.02	-0.60
Bilog (GAM)	-	-	-	-	0.07	0.57
NN reading level prediction	-	-	-	-	-0.01	-0.39
Average word length	-	-	-	-	0.10*	2.55
Item Format						
Extended Response	0.06	1.58	0.075	1.58	0.12	0.56
Multiple Choice	0.12**	4.74	0.15**	3.75	0.26	1.30
Short Response	0.03	1.26	0.04	1.34	0.10	0.49
Topic						
Earth and Space Science	0.03	1.30	0.05	1.28	0.12	0.60
Life Science	0.10**	4.74	0.12**	3.48	0.18	0.89
Physical Science	0.08**	3.42	0.095*	2.86	0.18	0.84
Science Practice						
Identifying Science Principle	0.18**	6.29	0.20**	4.98	0.26	1.24
Using Science Principle	0.05*	2.57	0.07	2.24	0.14	0.68
Using Scientific Inquiry	-0.02	-0.72	0.004	-0.10	0.08	0.38

Note. Statistical significance is based on comparing with Student's *t* distribution with 126 degrees of freedom (left) and 122 degrees of freedom (right) (*$p < 0.05$; **$p <0$ .01).

Another approach to identifying important features would be to add each linguistic feature individually to the format and cognitive features and compare the resulting regression models. The results are reported in Table 4.8 for these cases, the full feature set, and the case with no linguistic features. The best AIC, BIC and adjusted R^2 are obtained with average word length and all linguistic features. The linear regression model using average word length as the only linguistic feature gives the best result for AIC and BIC, as well as the best CV test R^2.

However, the CV R^2 measured for the average word length feature is not significantly better than other linguistic features in any training partitions and two of the test partitions.

Just as feature selection based on model fit to the full data may not generalize, selecting features based on correlation with the target variable using the full data set may not generalize. The linguistic complexity feature that has the highest correlation with *p*-value on the full data set (average number of word senses per word) is the most correlated feature for only one of the training splits and none of the test folds. The linguistic feature which leads to the best CV test R^2 (average word length) actually has a low correlation with *p*-value on the full set, but appears to be useful because of its correlation with cognitive demand features. Further, other linguistic features are more useful if we consider a different form of regression.

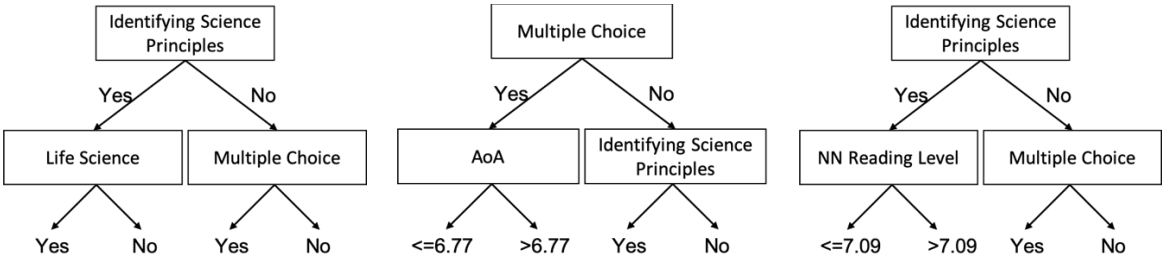
Table 4.8
Model selection criteria for LR and DTs with different linguistic features combined with format and cognitive features

Model	Linguistic feature	R^2	Adj R^2	AIC	BIC	Test R^2
Linear Regression	Age-of-acquisition	0.30	0.26	-71.34	-48.28	0.17
	Bilog (GAM)	0.30	0.26	-71.12	-48.06	0.17
	NN reading level	0.30	0.26	-71.12	-48.05	0.19
	Average word length	0.33	0.29	**-75.85**	**-52.79**	**0.21**
	All	**0.35**	**0.29**	-74.39	-42.68	0.20
	None	0.30	0.27	-72.96	-52.78	0.20
Decision Tree w/ DS	All	**0.41**	**0.39**	**-98.62**	**-84.21**	0.28
	None	0.38	0.36	-93.40	-79.00	**0.30**

Note. The best performance is highlighted in bold. The best performing models on all data do not perform the best on held out test data.

Since the best performing model on the test data from section Comparison of Model Selection Methods is the decision tree with depth selection, Table 4.8 also includes results for decision trees with all four linguistic features and without any. Decision trees are also often used

for identifying features that are important for prediction. The different trees that are learned in our experiments with depth selection are shown in Figure 4.2. The tree structure learned without linguistic features is shown 2(a); it is the same for each CV split. The features chosen are those that also have the highest significance in linear regression. When the AoA feature is included, the same tree structure is learned with the exception of one CV fold, which is illustrated in 2(b). Since the AoA feature is only used in one fold, the performance improvement that it brings is not significant. The tree structure learned using all linguistic features is shown in 2(c), and it replaces the topic feature with the NN reading level linguistic feature. When given all features, the tree chooses the NN reading level feature over average word length, since it can take advantage of non-linear dependencies between the input features and the target variable, leading to better CV test R^2 than all linear regression models. However, that benefit is not significant, and the best result is obtained when using no linguistic features. The two features that are consistently found to be important are the practice of identifying science principles and the multiple choice format, both of which tend to be easier for students.



(a) No linguistic features (b) AoA, CV split 2 (c) All linguistic features

Figure 4.2 Decision trees learned when given different feature sets.

Within a feature category (format, topic, practice), the binary features are not statistically independent, so it does not make sense to compare the significance of their *B* coefficients. However, we can understand the relative difficulty of the different cases in each category and the

significance of the *B* coefficients by using random sampling of the training set. When training the regression function using CV, each training partition gives different values of *B* because of random sample differences. If we generate more random samples (200 times, sampling 2/3 of the full data set), then we can examine the distribution of the coefficients. Figure 4.3 shows a violin plot, which is a visualization of the mean and variation of the linear regression coefficients learned from the 200 random samples. The biggest differences are in the science practices category, where identifying science principles is the easiest skill and using scientific inquiry is the most difficult. Among the topics, earth and space science is most difficult, and the multiple choice format is easiest. The factors that have coefficient distributions furthest from zero are the most significant for linear regression.

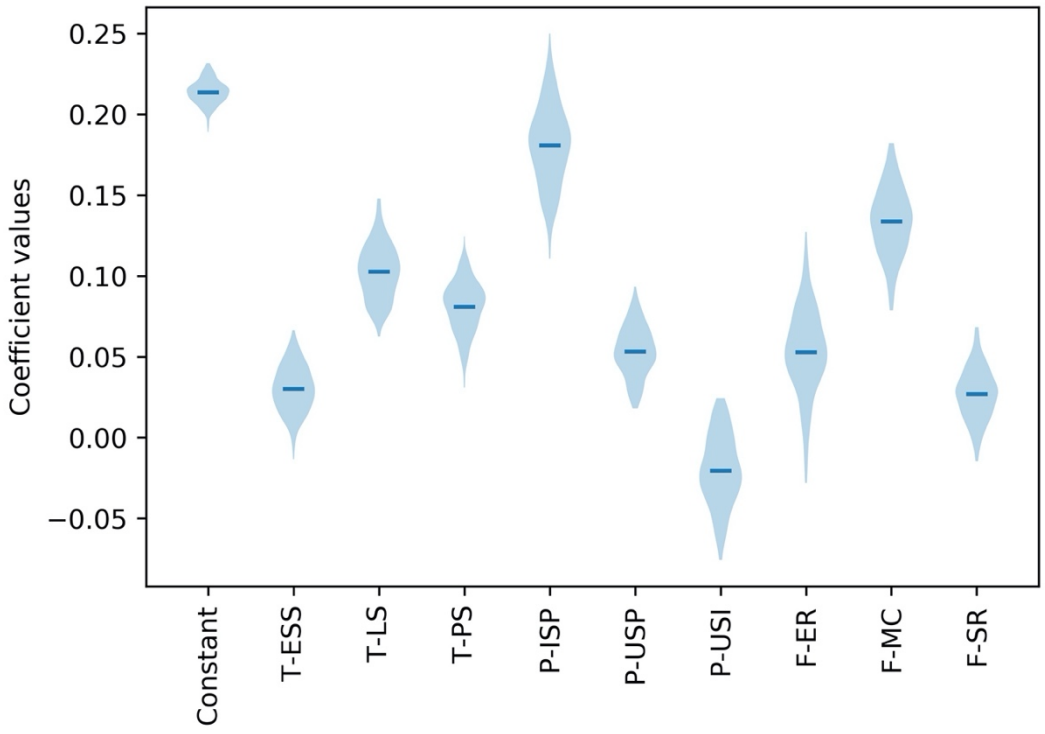


Figure 4.3 Violin plot for linear regression coefficients for 200 models trained to predict item *p*-values using 88 randomly sampled items for each model. The horizontal bar for each feature indicates the mean value of the coefficient. T-ESS, T-LS, T-PS refer to three topics: earth and space science, life science, physical science. P-ISP, P-USP, P-USI stand for science practices including identifying science principles, using science principles, using scientific inquiry. F-ER, F-MC, F-SR are short for extended response, multiple choice and short response (item format).

In summary, the experiments in this section illustrate how assessing feature significance when using the full data set can give misleading results due to overfitting. The same is true when choosing between features based on correlation with the *p*-values using the full data set. In addition, when input variables are not statistically independent or not linearly related to the target variable, linear regression analysis of significance can be problematic.

Training multiple regression models from random samples of the training set provides another method for exploring how item characteristics impact difficulty. Decision tree learning provides an alternative method for identifying important features. As an example, we show that a feature identified as useful in experiments on reading level prediction is more useful than both the best feature identified in linear regression analysis and the feature with the highest correlation with the *p*-values. These findings underscore the need to identify features using tasks and data other than items being used in the study. Lastly, through experimenting with multiple models and leveraging cross-validation, we confirm the utility of format and cognitive features in predicting item difficultly, but do not find significant effects from linguistic complexity.

4.4.3 Challenges in Interpreting Linguistic Features

As seen in the results presented above, using the average age-of-acquisition as an indicator of linguistic complexity does not improve item difficulty prediction, except for in the case of one instance of decision trees. The other linguistic features, word length and number of senses per word, seem to help in the case of linear regression, with improved performance on test data. For decision trees, however, having a linguistic feature, neural network (NN) reading level prediction, hurts the performance on held-out test data. This ties in partly with findings from El Masri et al. (2017) that linguistic features are not significantly predictive of item difficulty. Work by Höttecke

et al. (2018) indicates that linguistic features show an impact on item difficulty; however, it is not consistent across all items analyzed in the study. One cause of these findings may be that linguistic features are correlated with format and cognitive features, and thus it is difficult to see an impact when using linear regression. The second issue may be the length of items, with automatically extracted linguistic features not being useful for short texts. These hypotheses are explored next.

Interdependence of Linguistic and Cognitive Features. Here, we look for possible interdependence between the four linguistic features and either format or cognitive features by measuring the correlation between the average linguistic feature associated with a category of items (either response type, topic or practice) and the regression coefficient associated with the binary indicators for that category. In order to have several data points for computing the correlation, we use a method similar to the one used for generating the violin plot for coefficients: we randomly sample a subset of 88 items, compute the average value of each linguistic feature for each sub-category within the topic, practice and format categories, train a linear regression predictor with the format and cognitive features using this subset of data, and compute the correlation between each linguistic feature average and the regression coefficient associated with the corresponding indicators of either response type, topic or practice features. The results are presented in Table 4.9.

Table 4.9
Correlation values of linguistic features with linear regression coefficients of format and cognitive features for 200 linear regression models each trained on 88 randomly sampled items

Feature	Topic	Science Practice	Item Format
Age-of-acquisition	-0.55	0.11	0.17
NN reading level	0.57	-0.80	0.17
Bilog (GAM)	-0.48	0.04	0.15
Average word length	0.28	-0.65	-0.68

A positive regression coefficient corresponds to an easier category. For reference, life science is the easiest topic, identifying science principles is the easiest practice, and multiple choice is the easiest format. The age-of-acquisition feature seems to have an overlap with the topic regression coefficient, with a negative correlation since we expect items with a higher age of acquisition for words in a text would presumably be more difficult. The neural network reading level predictor has the greatest interaction with science practices, where we also saw the greatest differences between categories. Again, the correlation is negative, because higher reading levels should be more difficult. The average number of senses per word also has a strong positive correlated with science practices, not surprisingly because given its high correlation with item difficulty.

Overall, the neural network reading level has the highest correlation with cognitive features. Thus it is not surprising that it was chosen in the decision tree in place of the topic feature (Figure 4.2). Similarly the decision tree in Figure 4.2 using AoA instead of topic is consistent with the high correlation between the AoA and topic and low correlation with science practices. The results together show that interdependence of features pose a challenge for interpretation, and may in part explain existing mixed results on linguistic features.

Linguistic Features for Short vs. Long Texts. As claimed in El Masri et al. (2017), automatically extracted features do not work well for short texts. The work in Nadeem and Ostendorf (2018) showed on science texts that while a model using automatically extracted features works well for document length texts, the performance decreases drastically for texts shorter than 100 words. The explanation was that for shorter texts, the extracted feature vector tends to be sparse, with many more zero-valued features for short item texts compared to a longer document.

For the NAEP items in our study, the average item length is 69 words, and 86% of the items have less than 100 words. To see if items with longer contextual information can benefit from automatically extracted linguistic features, we repeat the *p*-value prediction study with 48 science items from PISA. The PISA items have average length of 216 words, and only 15% of the items have less than 100 words. Because there are so few PISA items, and because the NAEP and PISA topics are different, we limit the study to a simple linear regression using the features associated with item format, scientific practices, and age-of-acquisition. For a fair comparison, we reduce the number of NAEP items using multiple runs of training with a randomly sampled set of 48 items. Results for the entire selected set of items for both PISA and NAEP in Table 4.10. Adjusted R^2 and AIC are used rather than CV because of the very small sample size, and because these criteria gave selection results consistent with CV for format and cognitive features. We also include results with topic features for a comparison to the earlier results, which shows the performance degradation due to sampling (without age-of- acquisition, adjusted R^2 of 0.24 for the 48 items can be compared to 0.27 on the full NAEP data set). From the table, we see that for the longer PISA items, we get an improvement in adjusted R^2 using the age-of-acquisition linguistic feature, whereas there is a degradation for the shorter NAEP items. The AIC differences show the same patterns. This result supports the claim that item length impacts the potential utility of linguistic features, but work with a larger data set is needed to assess whether this generally holds.

Table 4.10
NAEP and PISA item difficulty prediction with and without the age-of-acquisition (AoA) linguistic feature

Item Source	Features	Adjusted R^2		AIC	
		w/o AoA	w/ AoA	w/o AoA	w/ AoA
NAEP	Format, Topic and Practice	0.25	0.23	-25.80	-23.80
	Format and Practice	0.27	0.25	-27.20	-25.20
PISA	Format and Practice	0.19	0.21	-27.60	-28.10

4.5 DISCUSSION

In summary, this work is aimed at providing insight into differences in findings of prior studies on factors that have been hypothesized to affect item difficulty. The goal is to identify characteristics of items that impact student performance and are likely to generalize to a new set of items. By providing detailed comparisons of model selection methods and new analyses of linguistic features, we demonstrate the utility of particular experimental methodologies and provide new findings related to the impact of item format, cognitive features and linguistic complexity on science assessment item difficulty.

When given a small amount of data, as is the case for work on science assessment items, a general problem with learning prediction models is overfitting. When using a large number of features or a model with more degrees of freedom (such as a decision tree), it is possible to fit one data set with high R^2, but have poor generalization to new data.

In the item assessment literature, standard methods to counteract for overfitting include adjusted R^2, AIC and BIC criteria, which provide a score that penalizes for the number of model parameters as a function of the amount of training data. In machine learning, a more common approach is to use a held out data set, or cross-validation when the available data is very limited.

Our experiments show that the different methods give similar results for simple models and a small number of features, but cross-validation is more stable across different types of models and feature sets. We also demonstrate that held out data (or CV) is useful for feature selection, either in decision trees with depth selection or other greedy selection techniques. Results for determining feature significance using linear regression show that it does not perform as well as decision tress. It is limited in its assumption of linear dependence with the target variable, as well an assumption that the predictors are independent, which we saw is not the case.

In using CV in model selection, our experiments on predicting item difficulty from item format, cognitive demand and linguistic complexity features provide support for findings that: i) response type is predictive of student performance in that multiple choice items are easier than constructed response, ii) both topic and scientific practices are associated with student performance. Like many others, we did not find a significant effect for linguistic complexity. However, we did find that linguistic complexity is correlated with other features that are significant, and it is difficult to decouple these factors, particularly when using linear regression. In addition, we provided results that support the hypothesis of El Masri et al. (2017) that linguistic complexity is difficult to characterize in short items. A reason for this is that word-level feature counts often used in characterizing linguistic complexity tend to be sparse – count-based features are higher in longer texts.

For future investigation, this work could be extended in two possible ways. First, we would like to obtain individual student responses for the released assessment items used in this study along with student demographic data. With individual student responses, we will be able to investigate how some of these factors might interact with demographic differences among students. More generally, our work in predicting difficulty of science assessment items is

constrained by the limited availability of public assessment items. In order to better understand how varying item characteristics make items difficult for students, the field would benefit from having more publicly available science assessment items with student performance data, including both the text and any graphics associated with specific problems. Having additional data would enable work in another direction, leveraging machine learning to automatically discover sources of item difficulty.

Chapter 5
Conclusion

Despite the extensive research on contextualized items over decades, I find it is still hard to claim we have fully understood the impact and consequences of using contextualized items. In fact, the deeper I dive into this topic, the little I feel I know about it. As De Lange (2007) points out, "the influence of contexts should be studied much more systematically than is presently the case, and we researchers should refrain from strong statements that we have proven to be of disputable quality until we have firmer evidence" (p. 1120). This dissertation serves as an attempt to clarify some common assumptions about six widely documented item characteristics and collect more evidence about the relationship between item characteristics and assessment outcomes. Using three methodological frameworks (IRT, CDM and machine learning), this dissertation unveils the differential impacts of the six item characteristics, which may be positive and/or negative, on students' performance and/or item statistics. Assembling the results from three studies, this dissertation delivers implications for assessment development, typically item writing.

5.1.1 Practical Implications

The first study in Chapter 2 addresses a common myth about contextualized items—richer contexts are better. By examining three levels of richness (abstract, contextualized, contextualized with illustration), the results of the study confirmed the positive effect of incorporating contexts to an abstract item while contradicted the assumption that adding illustrations to make the context even richer would bring additional benefits. These findings have important implications: while rich contexts bring benefits, it should be not abused. Prior studies

showed inappropriate contexts may introduce construct-irrelevant variances such as linguistic complexity (Ahmed & Pollitt, 2007) and misinterpretation of visuals (Ahmed & Pollitt, 2000). Item writers need to be aware of the tradeoff between the benefits brought by contexts and the potential risks posed by it. Besides, as an interaction effect was detected between topic and three levels of context richness, the results alarms the test writers to be cautious when developing contextualized items for different topics or even subjects, as some item characteristics may be more appropriate to be used with a certain topic.

Chapter 3 of this dissertation investigates the influence of context familiarity, indexed by daily life experience and classroom experience, on students' mastery profiles of cognitive attributes and two item parameters estimated from CDM. Based on the findings, if a student has been exposed to a context in the classroom that is similar to the context of the test item, the probability of he/she mastering the attributes required by the item may be impacted, positively or negatively. The study also revealed a potential effect of context familiarity (gained from classroom experience) on students' probability of answering an item correctly without knowing any required attributes. The findings of this study highlight the need for a careful examination of potential bias in item contexts. As this study identifies one potential source of bias (familiarity with an item context gained from classroom experience), there are many other factors that may lead to differential performance and thus should be carefully considered, such as gender and culture (McCullough, 2004). This discussion brings in another practical implication about how to detect bias in item contexts. The second study used students' self-reported survey questions as a proxy of context familiarity. More efforts should be made to develop assessment procedures that collects information that could be used to measure and control the impact of biased contexts in addition to students' responses.

Lastly, Chapter 4 explores the impacts of three item characteristics—cognitive demands, item format and linguistic complexity—on item difficulty. Cognitive demands are characterized by the physics topics the item taps and the science practices required by the item. Consistent with previous work, the study found item format and cognitive demands had a significant influence on item difficulty, while linguistic features showed no significant impact. The further investigation of linguistic features detected potential reasons for the minimal effect of linguistic complexity on item difficulty, which may be related to the high correlation between linguistic features and other predictors or the constraint of the short length of assessment items used in this study. The implication of this study points to the use of item characteristics in combination. When developing items, item writers could consider different combinations of item characteristics to adjust item difficulty. For example, items tapping a difficult topic or science practice could be presented in the format of multiple choice. Presenting an item at an appropriate level of difficulty is important for gauging students' true ability and discriminating students at various competency levels.

5.1.2 Limitations and Future Directions

This dissertation has several limitations that should be noted. First, all three studies suffer from the small sample size more or less, including both the number of participants and the number of items. The study in Chapter 3 contains only 16 items, which leads to the lack of power when detecting significant effects in regressions. The limited number of items ($N = 132$) in the study in Chapter 4 constrained the number of predictors that could be tested (for example, we did not test for the interaction among variables) and impacted the implementation of cross-validation. Another limitation across the three studies is that the reason for why a certain item characteristic affects students' performance or item statistics in a particular way has to be

inferred. For instance, in Chapter 2 we provided potential explanations for the negative impact of using contextualized-illustrated items based on previous literature. However, the proposed reasons could not be validated. The third limitation is specifically related to the application of CDM in the Chapter 3. The simulation study indicates an unsuccessful parameter recovery, which may be attributed to the misspecification of Q-matrix. Therefore, the reliability of the results needs further investigation.

To address those limitations, the future studies will be conducted with a larger research samples (both students and items) with a mix-method approach. In addition to statistical analysis, the incorporate of cognitive interviews with students will provide strong evidence for the results detected in the studies. Besides the increase in samples and inclusion of qualitative methods, I also would like to continue exploring findings that are not fully elaborated in the present dissertation for each of the three studies. More specifically, as an extension of the first study, I would like to dive into the interaction of topic and item characteristics, especially the use of visual representations. The obtained results may help test developers revise the blueprint of item development for different topics. For the second study, the focus of future work relates to the application of CDM. Chapter 3 demonstrates an effort to retrofit an assessment with a CDM. The short-term goal of future work is to revise the existing Q-matrix and replicate the current study. However, in the long term, I am interested in developing an assessment that is designed for the diagnostic purpose and examining whether the obtained results could be replicated. Besides, I am also interested in exploring new or existing models that incorporate item characteristics directly into item response function as covariates. In the third study we present a pilot study on the effect of linguistic complexity with items that have longer texts (PISA items). In the future work we

will continue this line of research by requesting more PISA items from the secured source and applying cross validation to understand the true effect of linguistic complexity on item difficulty.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219-234.

Ahmed, A. & Pollitt, A. (2000). Observing context in action. Paper presented at International Association for Educational Assessment conference. May14-19, Jerusalem.

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: An experimental investigation of focus. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 201-232.

Almuna Salgado, F. (2010). *Investigating the effect of item - context on students' performance on mathematics items*. Master's thesis. Available from University of Melbourne's Catalogue (melb.b4103780)

Almuna Salgado, F. (2017). The role of context and context familiarity on mathematics problems. *Revista Latinoamericana De Investigación En Matemática Educativa, 20*(3), 265-292.

Almuna Salgado, F., & Stacey, K. (2014). Item context factors affecting students' performance on mathematics items. In J. Anderson, M. Cavanagh & A. Prescott (Eds.), *Proceedings of the 37th annual conference of the Mathematics Education Research Group of Australasia (pp. 55-62)*. Sydney: MERGA. ISBN: 978-1-920846-27-5

Anderson, R. C. (1977). *Schema-directed processes in language comprehension. Technical Report No. 50*. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/17946/ctrstreadtechrepv01977i00050_opt.pdf?sequence=1&isAllowed=y

Anderson, R., Reynolds, R., Schallert, D., & Goetz, E. (1977). Frameworks for comprehending discourse. *American Educational Research Journal, 14*(4), 367-381.

Baldwin, S. G. (2008). *A bayesian testlet response model with covariates: A simulation study and two applications* (Order No. 3315471). Available from ProQuest Dissertations & Theses Global. (304567924). Retrieved from https://search.proquest.com/docview/304567924?accountid=14784

Barmby, P., Kind, P., & Jones, K. (2008) Examining changing attitudes in secondary school science. *Int Sci Edu*, 30, 1075-1093.

Bayraktar, S. (2009). Misconceptions of Turkish pre-service teachers about force and motion. *International Journal of Science and Mathematics Education*, *7*(2), 273-291.

Berends, I. E., & van Lieshout, E. C. D. M. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. *Learning and Instruction*, *19*(4), 345-353.

Boaler, J. (1993). The role of contexts in the mathematics classroom: Do they make mathematics more "real"? *For the Learning of Mathematics*, *13*(2), 12-17.

Boaler, J. (1994). When do girls prefer football to fashion? An analysis of female underachievement in relation to "realistic" mathematics context. *British Educational Research Journal*, *20*(5), 551-64.

Boonen, A., van Wesel, F., Jolles, J., & van der Schoot, M. (2014). The role of visual representation type, spatial ability, and reading comprehension in word problem solving: An item-level analysis in elementary school children. *International Journal of Educational Research*, *68*, 15-26.

Booth, J. L., & Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. *British Journal of Educational Psychology*, *82*(3), 492-511.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman & Hall.

Brookshire, J., Scharff, L., & Moses, L. (2002). The influence of illustrations on children's book preferences and comprehension. *Reading Psychology*, *23*(4), 323-339.

Brownell, W. A., & Stretch, L. B. (1931). *The effect of unfamiliar settings on problem-solving*. Durham, N.C: Duke University Press.

Caldwell, J. H., & Goldin, G. A. (1987). Variables affecting word problem difficulty in secondary school mathematics. *Journal for Research in Mathematics Education, 18*(3), 187-196.

Chen, H., & Chen, J. (2016a). Retrofitting non-cognitive-diagnostic reading assessment under the Generalized DINA Model Framework. *Language Assessment Quarterly*, *13*(3), 218-230.

Chen, H., & Chen, J. (2016b). Exploring reading comprehension skill relationships through the G-DINA Model. *Educational Psychology*, *36*(6), 1049-1064.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123-140.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Chen, Y., MacDonald G., & Leu, Y. (2011). Validating cognitive sources of mathematics item difficulty: Application of the LLTM to fraction conceptual items. *The International Journal of Educational and Psychological Assessment, 7*(2), 74-93.

Cho, S. (2016). *An application of diagnostic modeling to a situational judgment test assessing emotional intelligence* (Doctoral dissertation, University of Illinois at Urbana-Champaign, IL). Retrieved from http://hdl.handle.net/2142/95547

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). *Think you have solved question answering? Try ARC, the AI2 reasoning challenge*. arXiv preprint arXiv:1803.05457.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50*(1), 66-71.

Cook, M. (2006). Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education*, *90*(6), 1073-1091.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Common ideas about force and motion (n.d.). Retrieved from http://www.sci.sdsu.edu/crmse/nextgenpet/physics412/files/TL_Extension_J_PrepforL6_W15.pdf

Cortese, M. J., & Khanna, M. M. (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, *40* (3), 791–794.

Crisp, V. (2011). Exploring features that affect the difficulty and functioning of science exam questions for those with reading difficulties. *Irish Educational Studies*, *30*(3), 323-343.

Crisp, V., & Grayson, R. (2013). Modelling question difficulty in an A level physics examination. *Research Papers in Education*, *28* (3), 346–372.

Crisp, V., & Sweiry, E., (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 48*(2), 139–154.

Cui, Y., Gierl, M., & Chang, H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19-38.

de Ayala, R. (2009). *The theory and practice of item response theory (Methodology in the social sciences)*. New York: Guilford Press.

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), 343-362.

de la Torre, J. (2011). The Generalized DINA model framework. *Psychometrika*, *76*(2), 179-199.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development, Advance online publishing. doi:10.1177/0748175615569110*

De Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics* (pp. 1111-1142). Charlotte, NC: Information Age Publishing Inc.

Dong, D., Li, M., Ruiz-Primo, M., Zhai, X., & Minstrell, J. (2018). *Does context matter: Causal impact of exam problem context on student performance*. Paper Presented at the American Educational Research Association (AERA) Annual Meeting, New York, NY.

El Masri, Y., Ferrara, S., Foltz, P., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: The case of Key Stage 2 assessments. *Curriculum Journal*, *28*(01), 59-82.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129–157). Routledge.

Ferrara, S., & Duncan, T. (2011). Comparing science achievement constructs: Targeted and achieved. *Educational Forum*, *75*(2), 143–156.

Ferrara, S., Steedle, J. T., & Frantz, R. S. (2018). *Item response demands, predicting item difficulty, and validity of inferences from test scores*. Paper presented at the annual meeting of the national council on measurement in education, New York, NY.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359-374.

Fox, J. P., & Glas, C. A. (2016). Multilevel response models with covariates and multiple groups. In *Handbook of Item Response Theory* (Vol. 1, pp. 407-419). CRC Press.

Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly*, *18*, 277–294.

Gayawan, E., & Ipinyomi, R. A. (2009). A comparison of Akaike, Schwarz and R square criteria for model selection using some fertility models. *Australian Journal of Basic and Applied Sciences*, *3*(4), 3524–3530.

Garcia, M. (2014). Assessment practices past and future: Alternative approaches and teacher perceptions. *Honors Projects in Applied Psychology*. Paper 5. http://digitalcommons.bryant.edu/honors_appliedpsychology/5

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix. *Educational Researcher*, *40*(5), 223-234.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321. doi:10.1111/j.1745-3984.1989. tb00336.x

Haladyna, T. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn and Bacon.

Heller, P., & Hollabaugh, M. (1992). Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups. *American Journal of Physics, 60*(7), 637-644.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *Physics Teacher, 30*(3), 141-58.

Höttecke, D., Feser, M. S., Heine, L., & Ehmke, T. (2018). Do linguistic features influence item difficulty in physics assessments? *Science Education Review Letters*, *2018*, 1–6.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion model application to "Language" Assessment. *Language Testing, 26*(1), 31-73.

Javidanmehr, Z., & Sarab, M. (2017). Cognitive diagnostic assessment: Issues and considerations. *International Journal of Language Testing, 7*(2), 73-98.

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*(2), 186-203.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79-93.

Kane, M. (2006). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 17-64). New York: American Council on Education, Macmillan Publishing.

Khot, T., Balasubramanian, N., Gribkoff, E., Sabharwal, A., Clark, P., & Etzioni, O. (2015, September). Exploring Markov logic networks for question answering. In Proceedings *of the 2015 conference on empirical methods in natural language processing* (pp. 685–694). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D15-1080

Kirsh, D. (2009). Problem solving and situated cognition. In, P. Robbins & M. Aydede (Eds), The *Cambridge handbook of situated cognition* (pp 264-306). Cambridge: Cambridge University Press.

Klassen, S. (2006). Contextual assessment in science education: Background, issues, and policy. *Science Education, 90*(5), 820–851.

Kruger, C. (1990). Some primary teachers' ideas about energy. *Physics Education*, *25*(2), 86-91.

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: Example of PISA science items. *International Journal of Science Education*, *39*(4), 468–487.

Lee, J. (2011). *Second language reading topic familiarity and test score: Test-taking strategies for multiple-choice comprehension questions*, ProQuest Dissertations and Theses.

Lewalter, D. (2003). Cognitive strategies for learning from static and dynamic visuals. *Learning and Instruction, 13*(2), 177-189.

Li, M., Ruiz-Primo, M., Giamellaro, M., Wills, K., Mason, H., & Lan, M. (2012a). *Instructional sensitivity and transfer of learning at different distances: close, proximal and distal*

assessment items. Paper Presented at the American Educational Research Association (AERA) Annual Meeting, Vancouver, Canada.

Li, M., Ruiz-Primo, M., Wills, K., & Giamellaro M. (2012b). *Instructionally sensitive assessments across science modules*. Paper Presented at the American Educational Research Association (AERA) Annual Meeting, Vancouver, Canada.

Lim, Y. (2015). *Cognitive Diagnostic model comparisons* (Doctoral dissertation). Retrieved from http://hdl.handle.net/1853/53513.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253-275.

Ma, W., de la Torre, J., & Sorrel M. (2018). *The Generalized DINA model framework*. R package version 2.2.0. https://cran.r-project.org/web/packages/GDINA/GDINA.pdf

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*(3), 200-217.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*(3-4), 160-179.

Mayer, R. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*(2), 125-139.

Mayer, R., Sims, V., & Levin, Joel R. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology*, *86*(3), 389-401.

McCullough, L. (2004). Gender, Context, and Physics Assessment. *Journal of International Women's Studies, 5*(4), 20-30.

Mcmartin, F., Mckenna, A., & Youssefi, K. (2000). Scenario assignments as assessment tools for undergraduate engineering education. *IEEE Transactions on Education*, *43*(2), 111-119.

Meltzer, D. E., & Manivannan, K. (1996). Promoting interactivity in physics lecture classes. *The Physics Teacher, 34*(2), 72-76

Mesic, V. (2011). Modeling the discrimination power of physics items. *European Journal of Physics Education*, *2*(3), 5-19.

Mesic, V., & Muratovic, H. (2011). Identifying predictors of physics item difficulty: A linear regression approach. *Physical Review Special Topics-Physics Education Research, 7*(1), 010110-1– 010110-15.Mevarech, Z. R., & Stern, E. (1997). Interaction between knowledge and contexts on understanding abstract mathematical concepts. *Journal of Experimental Child Psychology*, *65*(1), 68-95.

Milenković, D., Segedinac, M., Hrin, T., & Gajić, G. (2016). Evaluation of context-level effect on students' performance and perceived cognitive load in chemistry problem-solving tasks. *Croatian Journal of Education, 17*(4), 959-982.

Morrison, K. M., & Embretson, S. E. (2014). Using cognitive complexity to measure the psychometric properties of mathematics assessment items. *Multivariate Behavioral Research, 49*(3), 292-293.

Nadeem, F., & Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 45-55).

National science education standards (1996). National Academy Press: Washington, DC. Retrieved from https://www.nap.edu/read/4962/chapter/7

Nijlen, D. V., & Janssen, R. (2015). Examinee non-effort on contextualized and non-contextualized mathematics items in large-scale assessments. *Applied Measurement In Education*, *28*(1), 68-84.

Paivio, A. (1990). *Mental representations: A dual coding approach* (Oxford psychology series; no. 9). New York; Oxford [England]: Oxford University Press; Clarendon Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12* (Oct), 2825-2830.

Perktold, J., Seabold, S., & Taylor, J. (n.d.). Statsmodels. *URL https://www. statsmodels. org/stable/index*.

Rosca, C. V. (2004). *What makes a science item difficult? A study of TIMSS -R items using regression and the Linear Logistic Test Model*. ProQuest Dissertations and Theses.

Ruiz-Primo, M., Li, M., Minstrell, J., Dong, D., Kanopka K., Hernandez, P., & Zhai, X. (2019a). *Contextualized science assessments: Addressing the use of information and generalization of inferences of students' performance*. Paper Presented at the American Educational Research Association (AERA) Annual Meeting, Toronto, Canada.

Ruiz-Primo, M., Li, M., Minstrell, J., Kanopka K., Hernandez, P., Dong, D., & Zhai, X. (2019b). *Testing the generalization to the domain inference: The use of contextualized clusters of items*. Paper Presented at the National Council on Measurement in Education (NCME) Annual Meeting, Toronto, Canada.

Ruiz-Primo, M. A., & Li, M. (2012, July). *The role of context in science items and its relation to students' performance*. Paper presented in the International Test Commission Bi-Annual Conference. Amsterdam, The Netherlands.

Ruiz-Primo, M., & Li, M. (2015). The Relationship between item context characteristics and student performance: The case of the 2006 and 2009 PISA science items. *Teachers College Record*, *117*(1), 1-36.

Ruiz-Primo, MA., & Li, M. (2016). PISA science contextualized items: the link between the cognitive demands and context characteristics of the items. *RELIEVE*, *22*(1), art. M11.

Rupp, A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Schnotz, W., & Bannert. M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*(2), 141-156.

Sheehan, K. M., Kostin, I., & Persky, H. (2006). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performance on the NAEP Grade 8 reading assessment*. Paper presented at the annual meeting of the national council on measurement in education, San Francisco, CA.

Sinharay, S., & Almond, R. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*(2), 239-257.

Solano-Flores, G., & Wang, C. (2015). Complexity of illustrations in PISA 2009 science items and its relationship to the performance of students from Shanghai-China, the United States, and Mexico. *Teachers College Record, 117*(1), 1-18.

Song, M., & Bruning, R. (2016). Exploring effects of background context familiarity and signaling on comprehension, recall, and cognitive Load. *Educational Psychology, 36*(4), 691-718.

Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *Journal of Early Adolescence, 37*(1), 85-128.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251-275.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

Turner, R., Dossey, J., Blum, W., & Niss, M. (2013). Using mathematical competencies to predict item difficulty in PISA: A MEG study. In *Research on PISA* (pp. 23-37). Springer.

Treagust, D. F., & Haslam, F., (1986). Evaluating secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier diagnostic instrument. (ERIC Document Reproduction Service No. ED383713)

Vajjala, S., & Meurers, D. (2014). Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, *165*(2), 194-222.

Valencia, S. W., Wixson, K. K., Ackerman, T., & Sanders, E. (2017). Identifying text-task-reader interactions related to item and block difficulty in the NAEP reading assessment. In

A publication of the NAEP Validity Studies Panel, San Mateo, CA: American Institutes for Research.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307. doi:10.1348/000711007X193957

Vos, V. (2014). *The Use of Context in Science Education*. Retrieved from http://dspace.library.uu.nl/bitstream/handle/1874/297294/The%20Use%20of%20Context%20in%20Science%20Education.pdf?sequence=2

Wang, C. (2012). *The use of illustrations in large-scale science assessment: A comparative study* (Doctoral dissertation, University of Colorado at Boulder, Boulder, CO). Retrieved from https://scholar.colorado.edu/educ_gradetds/26

Wang, T. (2016). *Examining sequence of contextualized items in science-experimental evidence on English learners (ELs) and Non-ELs* (Doctoral dissertation, University of Washington, Seattle, WA). Retrieved from http://hdl.handle.net/1773/35566

Wang, T., & Li, M. (2014). *Literature review of characteristics of science item contexts*. Paper presented at the Annual Meeting of National Association for Research in Science Teaching (NARST), Pittsburgh, PA.

Washburne, C. W. & Osborne, R. (1926a). Solving arithmetic problems. I. *The Elementary School Journal, 27*(3), 219-226. doi: https://doi.org/10.1086/461989

Washburne, C. W. & Osborne, R. (1926b). Solving arithmetic problems. II. *The Elementary School Journal, 27*(4), 296-304. doi:https://doi.org/10.1086/462005

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20* (1), 7-25.

Wiggins, G. (1993). Assessment: authenticity, context, and validity. *Phi Delta Kappan, 75*, 200. Retrieved from http://search.ebscohost.com.offcampus.lib.washington.edu/login.aspx?direct=true&db=eue&AN=503182444&site=ehost-live

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE-Life Sciences Education*, *15*(2), ar23.

Wainer H., Bradlow, E., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press: New York, NY.

Yamaguchi, K. & Okada, K. (2018). Comparison among cognitive diagnosticmodels for the TIMSS 2007 fourth grade mathematics assessment. *PLoS ONE*, *13*(2): e0188691. https://doi.org/10.1371/journal.pone.0188691

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Appendix A: Item Fit Statistics for Rasch Model

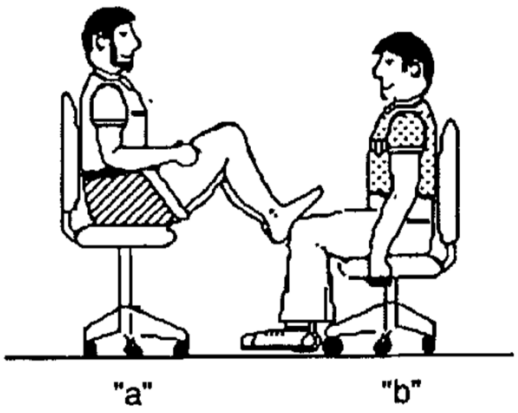
Item ID	Context	χ^2	Pr (> χ^2)
FM11	A	3.55	0.99
	C	1.54	1.00
	CI	0.58	1.00
FM12	A	7.12	0.76
	C	4.24	0.93
	CI	0.03	1.00
FM21	A	5.67	0.86
	C	15.71	0.14
FM22	A	14.03	0.19
	CI	25.14	0.02
FM31	A	0.41	1.00
	C	5.05	0.89
	CI	2.94	1.00
FM32	A	3.50	0.98
	C	1.45	0.99
	CI	2.17	1.00
FM41	A	3.69	0.98
	C	2.48	1.00
	CI	4.62	0.89
FM42	A	9.82	0.37
	C	3.59	0.97
	CI	5.57	0.93
E11	A	2.79	1.00
	C	3.82	1.00
	CI	1.39	1.00
E21	A	11.21	0.11
	CI	15.45	0.13
E31	A	2.51	1.00
	C	4.76	1.00
	CI	1.78	1.00
E41	A	16.88	0.12
	C	11.14	0.09

Appendix B. Results of Differential Item Functioning Analysis

Item ID	Context	Estimates	Adjusted *p*-value	R^2
FM11	A	4.81	0.51	0.04
	C	6.71	0.37	0.05
	CI	0.96	0.86	0.01
FM12	A	0.65	0.88	0.01
	C	2.14	0.78	0.03
	CI	1.33	0.83	0.02
FM21	A	0.25	0.91	0.00
	C	3.74	0.55	0.02
FM22	A	1.33	0.83	0.01
	CI	1.30	0.83	0.01
FM31	A	8.24	0.26	0.08
	C	10.79	0.15	0.10
	CI	0.40	0.91	0.00
FM32	A	3.96	0.55	0.05
	C	0.34	0.91	0.00
	CI	0.78	0.88	0.01
FM41	A	1.19	0.84	0.02
	C	2.23	0.78	0.02
	CI	3.11	0.67	0.03
FM42	A	1.59	0.83	0.02
	C	0.72	0.88	0.01
	CI	4.75	0.51	0.10
E11	A	0.59	0.88	0.00
	C	2.94	0.67	0.02
	CI	1.56	0.83	0.01
E21	A	1.51	0.83	0.01
	CI	0.27	0.91	0.00
E31	A	2.44	0.78	0.02
	C	4.71	0.51	0.04
	CI	4.12	0.55	0.03
E41	A	1.06	0.85	0.01
	C	0.10	0.95	0.00

Appendix C. Q11 in Force Concept Inventory

Two students, student “a” who has a mass of 95 kg and student “b” who has a mass of 77 kg sit in identical office chairs facing each other. Student “a” places his bare feet on student “b’s” knees, as shown below. Student “a” then suddenly pushes outward with his feet, causing both chairs to move.



11. In this situation,

(A) neither student exerts a force on the other.

(B) student “a” exerts a force on “b”, but “b” doesn’t exert any force on “a”.

(C) each student exerts a force on the other but “b” exerts the larger force.

(D) each student exerts a force on the other but “a” exerts the larger force.

(E) each student exerts the same amount of force on the other.

Appendix D. Selection of Linguistic Features

For characterizing linguistic complexity of items, we followed a multi-step process to identify a small number of candidate features using data sources other than the NAEP items, since there is a potential for bias when selecting features directly on the full data set. The approach was based on the assumption that features associated with higher grade-level science texts would be indicative of higher linguistic complexity in assessment items. We started with a set of 160 linguistic features that have been explored in work on automatic reading level prediction (Vajjala & Meurers, 2014). This set was narrowed down to 31 features by training a linear regression model with an L1 penalty to predict the grade level for a selection of passages from online open source K-12 textbooks[15]. We also learned non-linear functions of these 31 features by training a generalized additive model (GAM) using the same grade-level prediction task. A GAM learns a function of the form $\hat{y} = \sum_n f_n(x_n)$ where x_n are the features, and f_n the function learned for the *n*-th feature. The non-linear functions potentially make the features more useful than the original variables in linear regression. Additionally, we included a feature that was the grade level prediction of a neural network (also trained on the open source textbooks), which was shown to work better than feature-based models for short texts, including assessment items (Nadeem & Ostendorf, 2018). In the final step, we ranked the 63 features according to the correlation with the grade level of science items collected for grades 3-9 from online sources (Clark et al., 2018; Khot et al., 2015) (not including any NAEP items). In rank order, the selected four features with the highest correlation are:

- Average age-of-acquisition for all words in an item using word scores pulled from a psycholinguistic database (Cortese & Khanna, 2008);

[15] Michigan Open Book Project http://textbooks.wmisd.org/, Siyavula https://www.siyavula.com/

- Ratio of the log of unique words to the log of all words in an item (referred to as "bilog" in Vajjala and Meurers (2014)), transformed using a generalized function learned by training a GAM to predict grade level of text;
- Predicted grade level from a pretrained neural network model from Nadeem & Ostendorf (2018); and
- Average word length in characters.

Within the set of 63 items, the linguistic feature that had the highest correlation with the *p*-values of the NAEP items (average number of senses per word) did not have a particularly high correlation with the grade level of the online items. A problem with using the NAEP items to estimate correlation is that the estimates are unreliable because of the high variance associated with the small number of items. Specifically, we observe that the average number of senses per word did not have the highest correlation for any one of the three NAEP CV folds, and it was only best for one of the training splits. All three training splits were associated with different linguistic features in terms of correlation with *p*-values, and in no case did that feature have the highest correlation with the test fold. The high degree of variability in terms of which feature is best across different subsets of the data indicates that there is potential for overfitting. The five additional features that were selected by using the highest correlation with *p*-value for the three CV fold and training splits are:

- Non-linear transformation of sentence count learned by a generalized additive mode
- Non-linear transformation of count of *wh* words learned by a generalized additive mode, including what, where, who, how, whom etc.
- Average distance of main predicate from the start of sentence
- Average sentence length
- Average number of senses per word