Examining Sequence of Contextualized Items in Science-Experimental Evidence on English Learners (ELs) and Non-ELs

Ting Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Min Li, Chair

Catherine S. Taylor

Maria Araceli Ruiz-Primo

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Examining Sequence of Contextualized Items in Science - Experimental Evidence on English Learners (ELs) and Non-ELs

Ting Wang

Chair of the Supervisory Committee:

Associate Professor Min Li, Ph.D.

College of Education

Student performance on science tests is often substantially influenced by the context of individual test items. This study systematically describes the sequence patterns used in science test-item contexts and presents empirical evidence that demonstrate the ways in which the sequence patterns of contextual information support or hinder student performance on science tests. Contextual information includes supplemental information, such as a vignette or selected background facts that precede or follow a test item. This study addressed three gaps in the literature on contextualized items by providing the following in order to advance knowledge about how to achieve consistently equitable testing: a consistent theory-based framework for selecting test-item characteristics to be studied; new knowledge about how student performance is influenced by specific test item elements or features; and a better understanding of subgroups of students and their performance.

The design and data analyses were framed by three research questions: (1) How is providing the sequence of information presented in item contexts associated with student performance? (2)

Considering different levels of linguistic demands in various contexts, how is the contextual sequence linked to science performance of ELLs and non-ELLs? ESLs and non-ELLs? (3) How do different dimensions of sequence of context influence how students perceive and respond to tasks? This study included multiple facets of inquiry, including test-item creation, psychometrics, and student cognitive interviews. It generated item-development guidelines based on dimensions of context sequence. Items resulting from this process were designed to evoke students' stored knowledge relevant to the content and/or process skills being assessed. The newly developed science items were field tested with a diverse population of students from middle schools and high schools in China and the U.S. A range of psychometric and statistical procedures were applied to student test scores.

In particular, this project produced both theory and empirical evidence regarding how the sequence of contextual information can be attributed to differential performance among English as Second Language Learners (ESL), English language learners (ELL) and non-ELL groups. Findings from the ESL, ELL, and non-ELL comparisons led to important improvements in how items are developed to ensure test fairness.

Table of Contents

List of Figures

List of Tables

Dedication

I dedicate my dissertation work to my dear family. A special feeling of gratitude to my outrageously loving and supportive parents, Suxiang Wang and Shanshan Wang whose words of encouragement and push for tenacity ring in my ears. My parents encouraged and helped me at every stage of my personal and academic life, and longed to see this achievement come true. A dedication to my loving and amazing grandparents, Wenzhang Wang, Lingqie Zhou, and Jingtao Chen. I have been extremely fortunate in my life to have grandparents who have shown me unconditional love and support. The relationships and bonds that I have with my grandparents hold an enormous amount of meaning to me. My grandfather, Wenzhang Wang, passed away at the end of this year, but his support will always be with me.

I also dedicate this dissertation to my many friends who supported me throughout the process. I will always appreciate all they have done, especially Patricia Martinkova, Yuan-Ling Liaw, Kellie Wills, Phonraphee Thummaphan, and Ming-Chih Lan for helping me answer dissertation related questions, and Xiaoming Zhai, Duan Niu, Siwei Chen, Dongsheng Dong for helping with the field testing in China and the U.S.

Last, I dedicate this work and give special thanks to my best friends Yuan Zhang, Ying Lu, Huanshu Yuan, and Alec Kennedy for being there for me throughout the entire doctorate program. They have been my best cheerleaders.

Acknowledgements

I cannot express enough thanks to my committee for their continued support and encouragement: Dr. Min Li, my committee chair; Dr. Maria Araceli Ruiz-Primo; Dr. Catherine S. Taylor; and Dr. Adrian Dobra. I offer my sincere appreciation for the learning opportunities provided by my committee. A special thanks to Dr. Min Li, my advisor for her contributions of ideas and countless hours of reflecting, reading, and most of all patience throughout the entire process. The joy and enthusiasm she has for her research was contagious and motivational for me, and I thank her for supporting me to do the same.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. I would like to acknowledge and thank the College of Education from the University of Washington for awarding me the Doi doctoral research fund and assisting me to conduct my research. My work was also supported by the National Science Foundation, SES: 1461431, which enables a promising young scholar to establish a strong, independent research career.

I would like to thank the participating teachers, students, and administrators in middle schools and high schools from China and the Greater Seattle area who assisted me with this project. Their excitement and participation made the completion of this research an enjoyable experience.

Last, my time at the University of Washington was made enjoyable in large part due to the many friends and cohorts who became a part of my life. I am grateful for time spent with them, and for many other people and memories.

Chapter I: Introduction

1.1 Background of the Study

Item context is defined as the item component of supplemental information that precedes or follows a question in an item, such as a description of a lab setup, a natural phenomenon, or a practical problem (Ruiz-Primo & Li, 2012). The effect of real-world contexts on student test performance has been the subject of considerable research for several decades (Caldwell & Goldin, 1987; Clement, 1982; Little & Jones, 2010; Reisslein, Moreno, & Ozogul, 2010). Despite their widespread use in science testing, the utility and practice of contexts used in items have been called into question by some researchers (Ahmed & Pollitt, 2001; Wiliam, 1997). When test takers read and make sense of a contextualized item in science testing, certain characteristics of the context can evoke their stored knowledge during the process of interpreting what the question is asking and also activate their understanding of the scientific content and/or process skills. On the other hand, the context may adversely lead test takers to an incorrect path of generating improper problem-solving assumptions and therefore incorrect responses (Ahmed & Pollitt, 2001; Leighton & Gokiert, 2005).

Studying how characteristics of item contexts impact students' performance on tests directly addresses essential issues around the validity of score interpretations (Messick, 1993). When encountering contextualized items, students are expected to select and apply what they have learned to those particular contexts as evidence of being able to transfer their understanding. The context characteristics inevitably influence the test taking process, from reading the test item and interpreting what is being asked to selecting and applying the suitable knowledge and skills necessary to solve the problem (Ahmed & Pollitt, 2001). Therefore, score interpretations can be

directly or indirectly influenced by item contexts because of the important roles that contexts play in the test taking process.

Researchers have theorized item context characteristics and started to empirically study how some of them influence student performance. Researchers have focused primarily on math and reading (Boaler, 1993; Caldwell & Goldin, 1987; Clausen-May, 2006; Lehman & Schraw, 2002; Mevarech & Stern, 1997; Ozuru, Rowe, O'Reilly, & McNamara, 2008; Van Den Heuvel-Panhuizen, 2005), while there have been relatively few studies in science (Ahmed & Pollitt, 2001; Ruiz-Primo & Li, 2012). The research field has accumulated some understanding of how characteristics of item contexts influence student performance. For example, Caldwell and Goldin (1987), in their study of math word problems, found that abstract problems using symbolic objects were significantly more difficult than those with concrete situations using real objects such as hens and dogs. Yet this understanding is rather unsystematic and limited, especially in science assessment. In this dissertation, I focused on one particular characteristic of science item contexts, *the sequence of contextual information*, as no empirical study on contextualized items in science assessments has investigated this characteristic.

1.2 Purpose of Study

This study, Examining Sequence of Contextualized Items in Science - Experimental Evidence on English Language Learners (ELLs) and Non-ELLs, aims to examine how the characteristics of the sequence of contextual information presented in science items relate to student performance. Sequence of contextual information (or sequential cues of item contexts) refers to the order of descriptions involved in the context component of items, such as sequence

of events, sequence of intention and action, and sequence of cause and effect (Wang, Li, Thummaphan, & Ruiz-Primo, 2013).

Different patterns of sequential cues may require varying levels of cognitive sophistication for students to successfully decode the information presented, define and represent the underlying problem that needs to be solved, and subsequently respond to the item (Ahmed & Pollitt, 2001). I hypothesized that a well-organized sequence of information presented in the context should allow students to make sense and follow the logic of an item more easily, whereas items with a less organized sequence in the context may distract students from the scientific thinking being measured. Therefore, contextualized items with a well-organized sequence of contextual information are expected to result in better student performance.

My study also examined how linguistic demands of contextualized items impact student performance. Prior research on contextualized items has overlooked the effect of linguistic demands on student performance. Most studies fail to control for linguistic demands involved in different versions of contextualized items (such as abstract vs. concrete versions). For example, Caldwell and Goldin (1987) presented several pairs of math problems requiring computational skills in either an abstract or a concrete version. However, they did not examine the linguistic demands of either version used in their study.

Specifically, science tests are more linguistically demanding than math tests (Wolf & Leon, 2009). Related to this issue is how the linguistic demands of contexts may affect student performance, especially subgroups such as ESLs and/or ELLs. It has been argued that decoding of contexts presents a greater challenge for ELLs (Aguirre-Muñoz & Baker, 1998; Martiniello, 2009). Researchers have studied various item attributes (e.g., item length, syntactic and lexical

complexity) to identify possible patterns that might contribute to the underestimation of ELL performance on content assessments, such as math or science (Martiniello, 2009; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Solano-Flores, Barnett-Clarke, & Kachchaf, 2013; Wolf & Leon, 2009). They assert that some non-disciplinary language factors can threaten the validity and reliability of inferring students' understanding. For instance, minor changes in simplifying texts of test items can raise ELL students' test scores (Abedi, 2006; Abedi & Lord, 2001). However, no empirical studies have yet focused on the linguistic demands imposed by contextualized items to this type of students and how their performance is affected.

1.3 Research Questions

In this study, I examined the relationship between sequence patterns of contextual information and student performance after taking into account varying levels of linguistic demands. The research design and data analyses were guided by three research questions:

(1) How is providing the sequence of information presented in item contexts associated with student performance?

(2) Considering different levels of linguistic demands in contexts, how is contextual sequence linked to science performance of ELLs and non-ELLs? ESLs and non-ELLs?

(3) How do different dimensions of sequence of context influence how students perceive and respond to tasks?

In order to answer these three research questions, this study included two phases. Phase 1 focused on articulating the item development guidelines for dimensions of sequence. Collaboratively working with science content experts, I constructed science items with varying sequence patterns for each dimension while considering item linguistic demands. Phase 2

included both a field test of the developed science items with a diverse population of students from middle schools and high schools in China and the U.S., and cognitive interviews with a stratified sample of students from both China and the U.S. I then performed the psychometric and statistical analyses with test scores and coded the student interviews to technically evaluate the items and to examine how patterns of sequential cues in item contexts were associated with student test scores.

1.4 Significance of the Study

Educational reform efforts have stressed the importance of teaching and testing in meaningful contexts. As a result, contextualized items have become widely used in science testing. For example, 70% of the National Assessment of Educational Progress (NAEP) released science items for Grade 8 (71% for Grade 4) and 78% of 2011 the Trends in International Mathematics and Science Study (TIMSS) released science items for Grade 8 are contextualized (Wang & Li, 2014). Contextualized items are the primary item type in certain assessment programs, such as the PISA, AAAS science assessment, or science assessments used in the state of Washington. Unfortunately, contextualized items are developed based mainly on either conventional wisdom or on writing rules summarized from non-contextualized items. Contexts used in items may interfere with the target construct and lead to inaccurate inferences about student learning.

In this study, I systematically investigated contextualized items to examine the sequence information of item contexts, conducted student cognitive interviews to unpack student thinking processes, and advanced our knowledge about equitable testing. Those innovations resulted in two significant contributions.

First, this exploratory study advances understandings of the development of contextualized items in the science of item development and potentially in other Science, Technology, Engineering and Math (STEM) fields. The Next Generation Science Standards (NGSS, 2013) have shifted priorities to teaching science and engineering practices in conjunction with content, which brings an assessment challenge because current measurements and approaches do not allow these types of performances to be assessed easily (Coffey & Alberts, 2013). Moreover, Welch (2006) notes that there is an urgent need for research on item writing, as such items are the backbones of the assessment industry. This study contributes sound and solid methods to the assessment field by offering an empirically tested approach to unfolding and characterizing the sequential cues used in science item contexts. In addition, multiple sources of evidence were collected to validate the item manipulation and provide concrete recommendations around the item development that are applicable to classroom assessments and large-scale testing in the STEM education.

Second, the inclusion of linguistic demands along with sequence of contextual information in test items is critical to evaluate the item impact on student performance, especially sub-groups with limited English proficiency. Based on the 2015 National Center for Education Statistics Fact report, there were about 4.4 million ELLs in United States public schools in school year 2012-2013 (9.2 percent), which was higher than in 2002-2003 (8.7 percent). With the rapidly growing diverse K-12 student population in the U.S., this study contributes to the equitable measurement of all students (NGSS, 2013). By offering valuable insights with which to more fully consider item characteristics (i.e., sequence and linguistic demands of item contexts), my research work advances the ELL testing research. Beyond simply comparing score differences

between ELL and non-ELL groups, my findings may help educators and teachers to devise effective ways of assessing ELLs.

1.5 Organization of the Dissertation

I divided this dissertation into five chapters. Chapter I – Introduction describes the background and importance of the research questions raised in this study. Chapter II – Literature Review justifies reasons to use contextualized items in the STEM assessments, discusses the literature base of the context characteristics in contextualized items, and summarizes studies in the ELL testing. Chapter III – Methods describes the item development, participants, data collection, data coding, and analyses that I used in this dissertation. Chapter IV – Findings organizes and reports the study's main findings presented in the order of the research questions asked. Chapter V – Conclusions and Recommendations presents a set of concluding statements, along with implications from the findings and suggestions for future study.

Chapter II: Literature Review

Contextualizing items is an effective strategy to assess complex thinking and to determine the extent to which students are able to apply and transfer their knowledge (Ahmed & Pollit, 2007; Boaler, 1993; Haladyna, 1997). In this chapter, I first justify the definition of what item contexts are from the literature and then discuss reasons to use or not use contextualized items in STEM assessments. I then analyze the literature of context characteristics in contextualized items to synthesize what previous studies have done. Next, I review literature on ELL testing to identify possible linguistic sources in item contexts accounting for instability of student performance. Last, I state the focus of this study and discuss gaps in research around item contexts. This section demonstrates the need for further research in the area of science contextualized items for testing students with limited English proficiency.

2.1 What Are Item Contexts

Researchers define item contexts from three different perspectives: science teaching, assessments, and item generation. Item contexts used in science teaching are defined as a selected or created display consisting of information that is directly relevant to the intended learning outcomes and novel to students (NGSS, 2013). Item contexts used in assessment include characteristics of an item that are used to frame a question and the response choices (Kirsh, 2009). The notion of item contexts in item generation research is broad, referring to the collection of factors related to the way in which an item is presented, such as minor changes in wording and format, or the specific way a question is asked (Davey & Lee, 2011). Although item contexts are approached slightly differently depending on their various purposes, item

contexts include descriptions of science scenarios, lab setups, or practical problems (Ruiz-Rrimo & Li, 2012).

2.2 Why Use/Not Use Contextualized Items?

In this section, I discuss the merits and concerns of using contextualized items in science teaching, assessments, and item generation.

2.2.1 Merits of Using Contextualized Items

Contextualized items have been widely used in science teaching and testing as part of recent reform efforts. Its proponents often focus on three perspectives: contexts used in science teaching, assessments, and item generation (Wang & Li, 2014).

Contexts Used in Science Teaching. In alignment with the NGSS (2013), emphasis on context-based teaching aims to broaden and deepen the integrated understanding of science content and scientific processes. Three practical reasons are cited to support the use of contexts in science instruction. First, context-based problems and projects are used by teachers to motivate students in classroom learning (Taylor & Nolen, 1996). Based on prior cognitive studies, contextualized items may offer a more engaging background for students to solve the problem than items without context (Pollitt & Ahmed, 2000). Therefore, contextualized items are supposed to be perceived by students as something realistic, interesting, and relevant in contrast to typical textbook problems (Haladyna, 1997).

Second, some researchers (Barber, 2000; Bellocchi, King, & Ritchie, 2011; Hofstein, Kesner, & Ben-Zvi, 1999) have argued that contextualized instruction using societal issues and technological applications may stimulate student thinking in science learning. Such context

makes students feel more positive about science by solidifying the importance of what they are studying.

Third, other researchers (Rivet & Krajcik, 2008; Taber, 2003) have pointed out that context-based teaching invites students to use their prior knowledge and everyday experiences, which can function as catalysts for understanding challenging science concepts. Learning occurs when students process new information or knowledge in a way that makes sense to them in their own frames of reference. Therefore, contexts used in science teaching act as vehicles to help student learning occur both extrinsically and intrinsically (Beasley & Butler, 2002).

Contexts Used in Assessments. Contextualized items are considered an important assessment method in the field of STEM assessment given their wide use. We found that 28% of NAEP released algebra math items at Grade 8, 70% of NAEP released science items at Grade 8 (71% at Grade 4), 78% of 2011 TIMSS released science items at Grade 8, and all PISA released science items are contextualized items (Wang, Li, Thummaphan, & Ruiz-Primo, 2013).

Three reasons are discussed in the assessment literature for using contexts in assessment:

(1) Contexts in assessment may make items more relevant to students, as contexts provide a familiar setting in which students can recognize what they are asked to do (Crisp, 2011; Turmo & Elstad, 2009). In addition, contexts can make abstract concepts concrete by connecting student experience with formal scientific principles (Taber, 2003). Here, student experiences resulting from formal and informal learning can appear in social, cultural, physical, and psychological forms.

(2) Contextualized items could potentially assess learning skills involved in the process. Skills, such as understanding contextual clues (and constraints) when applying a particular

scientific law or engineering design rule, are valuable learning outcomes (Clausen-May, 2005; Nickson & Green, 1996). The way to adequately measure such skills is to include contexts (Perin, 2011). In other words, in the testing situation, contexts must be included as the stimulus in order to elicit desirable students' responses or behaviors with respect to the underlying construct – applying the skills that we would like to measure.

(3) Use of contexts allows the assessment of students' ability to apply what they have learned (i.e., transfer of learning) instead of merely assessing how well they have memorized the textbook content (Ahmed & Pollitt, 2007). Transfer of learning has been defined as "the ability to extend what has been learned in one context to a new context" (Bransford, Brown, & Cocking, 2000, p. 39). Thus, it is expected that contextualized items will require students to apply what they have learned in novel contexts (Ahmed & Pollitt, 2007). In other words, concepts and information in context project students into a practical applied world rather than a catechism of questions. For example, instead of asking students to write down the definition of optimum environment, test developers may present a scenario in an item asking for the optimum environment for barley. In that way, contextualized items can potentially assess the deep learning inherent in transfer of knowledge rather than that of rote memory used to learn isolated information from textbooks (Perin, 2011).

Contexts Used in Item Generation. In recent decades, several researchers have investigated the cognitive processes involved in solving contextualized items (Anderson, Reder, & Simon, 1999; Pollitt & Ahmed, 1999) and the item features that influence their difficulty (Enright, Morley, & Sheehan, 2002). Research has identified two benefits to using contexts in item generation: (1) Contexts present certain sets of scheme-based relationships that tend to occur in

the world. The problem solving of contextualized items requires the retrieval of the appropriate schema based on the comprehension of the contexts. Therefore, contexts can offer useful models that include the schema equations, variables, and other linguistic and numeric constraints for developing test items in a variety of ways (Singley & Bennett, 2002). (2) Contexts can potentially reduce construct irrelevance for item generation, and thereby help students focus on activating relevant schema for problem solving (Ahmed & Pollitt, 2001; Heredia, Furtak, & Morrison, 2012).

2.2.2 Concerns Raised about Using Contextualized Items

There are two concerns related to using contextualized items. First, developing items with contexts is demanding in terms of time and resources. Compared to developing items without contexts, item writers need to spend more time considering grade-level appropriate scenarios for students. Also, linguists need to carefully review the language used in item contexts to make sure the language has a high level of standardization as a cultural necessity (Garvin, 1973).

In a recent mathematics study, Hickendorff (2013) assessed the effects of contextual arithmetic problems on sixth graders' problem solving skills. Multi-digit arithmetic problems presented in two conditions – with and without a realistic context – were solved by sixth graders in the Netherlands. Hickendorff found that there were no positive effects from presenting arithmetic problems in a realistic context typical for school mathematics tests when testing addition, subtraction, and multiplication. In other words, contexts do not matter in those cases. Therefore, it is not worthwhile to spend extra time and energy developing contextualized items if contexts do not matter in measuring a particular set of student skills.

Second, presenting items in contexts may unintentionally generate test-irrelevant information, thereby threatening the score interpretation by involving contexts. Ahmed and Pollitt (2007) argued that contextualized items contained a great deal of irrelevant information, some of which may intentionally require students to select what is relevant to answer the question. Yet in exam conditions where students are under stress, the brain has limited attentional resources to deal with irrelevant information exterior to the science being tested by the question (Baddeley, 1986). For example, contexts that are unfamiliar or non-meaningful to students can exert debilitating effects on their ability to solve the problem (Mevarech & Stern, 1997). In such cases, students may not be able to identify the relationship between the contexts and the particular scientific principle embedded in the problem. In such cases, context brings irrelevant information that may prevent students from problem solving. Thus, it is recommended that irrelevant information be kept to a minimum (Ahmed & Pollitt, 2007).

In sum, studying how characteristics of item contexts impact students' performance on tests is important to address validity issues around score interpretations. When encountering contextualized items, students are expected to select and apply what they have learned with respect to particular contexts as evidence of being able to transfer what they have learned. Characteristics of contexts inevitably influence the test-taking process, ranging from reading the test item and interpreting what is being asked to selecting and applying the suitable knowledge and skills (Ahmed & Pollitt, 2001). In the next section, I focus on four specific characteristics of item contexts in relation to student performance or item difficulty parameters: abstract/concrete, focused/non-focused, pictorial/textual, and familiar/unfamiliar contexts.

2.3 Relevant Literature on Context Characteristics in Contextualized Items

In a recent research on context characteristics, Seiler (2012) theorized two desired characteristics of contexts, comparability and translatability. Comparability refers to the extent to which a particular context used in an item is as good as any alternative ones. Translatability indicates the extent to which context, rather than cause confusion or misinterpretation, allows test takers to recognize the content presented in an item, use the context to activate and integrate what has been learned, and apply scientific knowledge to answer the item. In other words, comparability is about the reliability of contexts in terms of their exchangeability. Research methods used in some studies have focused more on comparability than on translatability in the sense that student scores were compared between different versions of the same test. Translatability is about the validity of contexts, as it indicates whether or not contexts successfully serve their purposes. Studies on translatability use cognitive interviews or similar methods to determine the validity of contexts. This dissertation work is grounded in the translatability aspect of item contexts.

Empirical studies have identified at least four additional characteristics of contexts, all of which are associated with student performance on test items. Characteristics examined in prior research include abstract/concrete (e.g., Caldwell & Goldin, 1987), focused/non-focused (e.g., Ahmed & Pollitt, 2007), pictorial/textual (e.g., Ruiz-Primo & Li, 2012), and familiar/unfamiliar (e.g., Crisp, 2011).

Previous publications on item contexts were carefully collected from three sources: (1) University of Washington library electronic databases, using primary search by key words such as contexts, context-based, and contextualization; (2) literature network search as a secondary search via cited references or paper lists in annual conference presentations (via the electronic

program), such as National Association for Research in Science Teaching (NARST), and American Educational Research Association (AERA); and (3) expert recommendations via colleagues to reduce blind spots in the document collection.

Peer-reviewed papers and academic downloadable articles often yield a high-quality set of technical papers. The following criteria were used to select the articles synthesized in this study: (1) one or more characteristics of item contexts are discussed, although the authors may not have labeled items as contextualized items; (2) articles in STEM fields either include a theoretical discussion or provide empirical evidence regarding impacts of item contexts on item parameters or student performance; (3) studies are reported as a published article in journals or conference papers between 1982 and the present. The year of 1982 was chosen because the think-aloud protocol was first used in John Clement's 1982 pioneering investigation of the diagramming problem solving process in contextualized problems.

Five types of studies were collected that examined those characteristics: (1) *theoretical articles* that explain or illustrate the importance of item contexts, although no study was conducted to examine the context characteristics on item parameters (e.g., Ahmed & Pollitt, 1999; Boaler, 1993; Grawe, 2011; Taber, 2003); (2) *empirical studies – teaching and learning* in which characteristics of contexts were empirically studied from perspectives of effective teaching and learning (e.g., Beitzel, Staley, & DuBois, 2011; Kelly, 2007; Reisslein, Moreno, & Ozogul, 2010; Rivet & Krajcik, 2008); (3) *empirical studies – measurement*, including papers on assessment and on item generation, in which at least one characteristic of item contexts was examined in either comparative or non-comparative studies (e.g., Ahmed & Pollitt, 2001; Kaminski, Sloutsky, & Heckler, 2008); (4) *secondary data analysis papers* in which item

contexts were discussed by using existing large-scale tests, such as state tests or TIMSS (e.g., Gutierrez & Ikeda, 2009; Kastberg, D'Ambrosio, McDermott, & Saada, 2005; Schneider, Huff, Egan, Gaines, & Ferrara, 2013); (5) *qualitative studies* that employed only cognitive interviews or other qualitative analyses to investigate how contexts are linked with student performance (e.g., Gabora & Kitto, 2013; Roth & Hwang, 2006).

2.3.1 Abstract/Concrete Contexts.

Nine studies examined or discussed abstract and concrete characteristics, although they approached it with considerably different perspectives (see Table 1). In one secondary data analysis study of contextualized items in science, Ruiz-Primo and Li (2012) conceptualized the characteristic as the degree to which explanations or specific events were able to make the underlying abstract ideas more tangible and accessible.

In contrast, other researchers in the mathematics domain conceptualized this characteristic as dichotomous. Caldwell and Goldin (1987) defined a problem statement describing only abstract or symbolic objects as abstract (e.g., left item in Figure 1), and a problem statement describing a real situation dealing with real objects as concrete (e.g., right item in Figure 1). Similarly, in a recent study on electrical engineering, Moreno, Ozogul, and Reisslein (2011) applied this dichotomous notion of abstract and concrete to operationalize the visual representation in addition to the problem statement. Mevarech and Stern (1997) defined abstract visual representations as those that use conventional symbols to represent the relevant elements of a problem (e.g., left item in Figure 2), while concrete visual representations were defined as those that illustrate the real-life objects corresponding to a problem (e.g., right item in Figure 2).

Based on the literature, I define concrete contexts as those involving facts and descriptions about real-life objects, and abstract contexts as those focusing more on expression of conceptual ideas.

Table 1

Number of Papers Examining Abstract vs. Concrete Characteristics

Type \ Domain	Engineering	Mathematics	Science
theoretical articles	0	1	0
empirical studies – teaching and learning	2	1	0
empirical studies – measurement	0	2	0
secondary data analysis	0	0	2
qualitative studies	0	1	0

The research on the abstract/concrete characteristic in contextualized items is still in its fledgling stage. This characteristic has been studied along with other characteristics (e.g., factual versus hypothetical). In several studies, findings on the abstract/concrete characteristic have been confounded with other characteristics. One such representative study by Caldwell and Goldin's (1987) examined the relative difficulties of abstract versus concrete word problems in mathematics for junior and senior high school students. Figure 1 shows a pair of example items: the abstract word problem is the problem in which the mathematical computation is explicitly described using numbers (e.g., a given number is six more than a second number), and the concrete word problem presents a similar computation by using real objects (e.g., a young farmer has eight more hens than dogs). Abstract and concrete word problems were created in a set. Several sets of problems were combined and tested by students at different grade levels.

Caldwell and Goldin reported that problems of abstract mathematical relations were significantly more difficult for students than those with concrete situations ($p < .01$) described in story problems. They speculated that concrete contexts could connect to students' prior

knowledge of similar word problems and enable them to use mathematical principles, thereby resulting in higher scores. Thus, problem solving can be facilitated when students experience concrete contexts. Their findings were consistent at both the junior and senior high school levels, but the differences became smaller in magnitude with increasing grade level. For example, junior high school students solved 55% of the concrete problems versus 43% of the abstract, while senior high school students solved 69% versus 66%, respectively. Caldwell and Goldin (1987) pointed out that these differences may be due to senior students possibly having developed formal operational thinking skills to comprehend and identify the underlying mathematical structure of a word problem. Abstract problems may no longer be as challenging for senior students because their thinking skills are improving and the scaffolding of concrete contexts is less necessary for them.

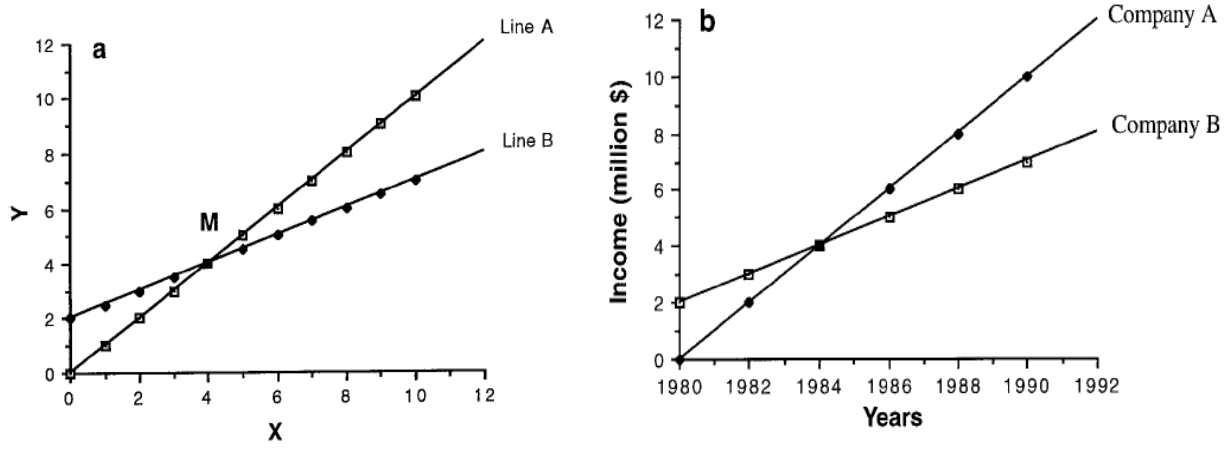
• The value of a given number is six more than the value of a second number. The sum of two times the first number and four times the second number is 126. What is the value of the second number?	• A young farmer has eight more hens than dogs. Since hens have two legs each, but dogs have four legs each, all together the animals have 118 legs. How many dogs does the young farmer own?

Figure 1. Two examples from Caldwell & Goldin, 1987, p.189.

Note. The left item is an abstract version; right item is a concrete version.

Mevarech and Stern (1997) investigated the effects of sparse versus real contexts on the understanding of slope when interpreting linear graphs (see Figure 2). Their notions of sparse and real contexts were aligned with the definitions of abstract and concrete used by Caldwell and Goldin (1987). Junior high school students were randomly assigned into one of the two versions of the problem. Mevarech and Stern reported that the mean score of the sparse context (Mean = 2.5) was significantly higher than the one embedded within real contexts (Mean = 2.1, $p < .05$), indicating the sparse context facilitated understanding of abstract mathematical concepts. They

speculated that math problem embedded in a sparse context activates abstract logic-mathematical knowledge structures, whereas the other version embedded in real contexts that lack fine-grained mathematical structures results in low student performance.



- Figure above shows a graph of two lines A and B that are intersecting at point M. Q1: Up to point M, is the growth rate of line A greater than, less than, or equal to the growth rate of line B? Please explain your reasoning.

- Figure shows a graph representing the incomes of company A and company B between the years 1980 and 1990. Q1: Until 1984, was the income growth rate (i.e., the growth in income per year) of company A greater than, less than, or equal to the income growth rate of company B? Please explain your reasoning.

Figure 2. Two examples from Mevarech & Stern, 1997, p. 75.

Note. The left item is a sparse version; right item is in a real context version.

The above two studies reach different conclusions regarding abstract versus concrete contexts on item difficulty parameters. Two reasons might explain the differences: First, the student sampling populations came from two different countries: the U.S. (Caldwell & Goldin, 1987) and Israel (Mevarech & Stern, 1997). Approaches to student learning and thinking about mathematics might be different in the two countries regarding the extent and amount of concrete examples teachers use and how much abstract reasoning is emphasized in the curriculum. Second, the mathematics domains tested in the two studies were different: one focused on a

mathematical computational algorithm (Caldwell & Goldin, 1987), while the other focused on interpreting line graphs (Mevarech & Stern, 1997). It is possible that the skills necessary to successfully solve the problems in each of these two mathematical domains are different.

Other studies examined this characteristic from the perspective of the instructional effect, for example, Moreno, Ozogul, and Reisslein's (2011) study of high school students and Flores' (2010) study with elementary school special needs children. Both studies examined the instructional effect of abstract/concrete contexts from the perspective of using visual representations. They did not examine how abstract/concrete characteristics of the test construct affect student performance. Therefore, these two studies of learning and teaching perspectives did not provide clear-cut evidence of the mechanism by which context characteristics impact student performance. Overall, results are inconclusive regarding the way in which abstract versus concrete contexts impact item parameters.

2.3.2 Focused/Non-focused Contexts.

The focus of a context is defined as the extent to which aspects of the context perceived by students that will be the most salient in real life (Ahmed & Pollitt, 2007). In light of Ahmed and Pollitt's ideas, Ruiz-Primo and Li (2012) consider a context to be focused if it activates critical underlying scientific concepts tapped in the question, and a context to be unfocused if it deviates students' attention to irrelevant aspects of the context that interfere with the comprehension and framing of the question. In the reviewed literature, there are two papers examining this characteristic in STEM fields (see Table 2).

Table 2

Number of Papers Examining Focusedness Characteristics

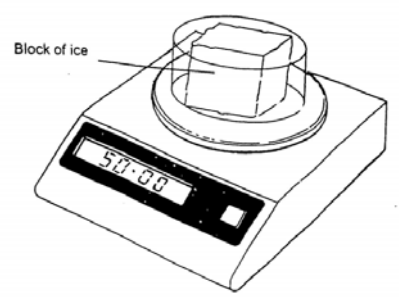
Type \ Domain	Engineering	Mathematics	Science
theoretical articles	0	0	0
empirical studies – teaching and learning	0	0	0
empirical studies – measurement	0	0	1
secondary data analysis	0	0	1
qualitative studies	0	0	0

In an empirical study, Ahmed and Pollitt (2007) examined whether the degree of focus of a science question's context can affect student performance. Items differing in their focusedness levels (e.g., non-focused vs. focused; non-focused, focused, and very focused; see example items in Figure 3) were constructed and assigned in twelve different forms via test booklets. The scientific idea behind the example item is tested by asking students to decide what happens to the mass of an object when it changes state. Student interviews revealed that using an ice lolly (i.e., popsicle) as an example added construct-irrelevant variances in that students were distracted by food wastage resulting from the melting ice lollies. The focus of this item could potentially be changed to waste and loss instead of state changes; thus, the item on the right in Figure 3 is less focused than the one on the left.

Ahmed and Pollitt (2007) concluded that questions with more focused contexts (see left item in Figure 3) proved easier for students than those with less focused ones (see right item in Figure 3, $p < .001$), as items with focused contexts better elicited the anticipated cognitive processes in students than did those with non-focused contexts. By focusing the item's context on the intended scientific concepts, construct-irrelevant difficulties are minimized. In contrast, unfocused contexts provoked unnecessary misunderstandings instead of eliciting what was

required to answer the question. In other words, unfocused contexts could cause distractions and could be a threat to the fidelity and validity of the construct (Ahmed & Pollitt, 2007). In sum, students performed better on items with focused contexts than on those with unfocused contexts.

- Peter was investigating what happens when ice melts. He put a block of ice on his electric scales. The scales showed a reading of 50g. This is shown in the diagram.



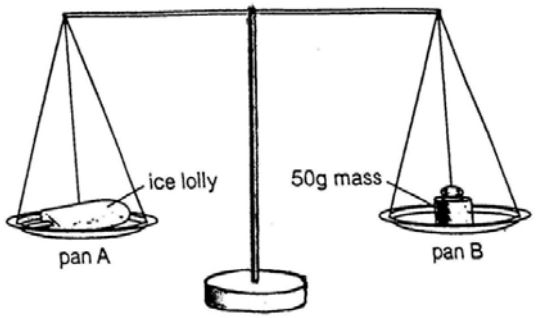
Peter left the scales like this for 15 minutes until the ice had melted. He then took a reading from the scales. Place a tick in the box that describes the scale reading after 15 minutes.

☐Greater than 50g

☐Less than 50g

☐50g

- Peter put an ice lolly on one pan of a balance. He balanced this with a 50 g mass on the other pan. This is shown in the diagram.



Peter left the balance like this for 15 minutes until the ice lolly had melted. Place a tick in the box that describes the balance after 15 minutes.

☐Pan A has moved up.

☐Pan A has moved down.

☐Pan A and pan B are still at the same level.

Figure 3. Two examples from Ahmed & Pollitt, 2007, p. 224-226.

Note. The left item is a very focused version; right item is an unfocused version.

2.3.3 Pictorial/Textual Contexts.

Pictorial contexts refer to contexts represented with nonlinguistic schematic representations such as pictures, graphs, symbols, or tables. Textual contexts refer to contexts that lack pictorial

representations and contain only linguistic material. In the reviewed literature, there are seven papers examining this characteristic in the STEM field (see Table 3).

Table 3

Number of Papers Examining Resources Characteristics

Type \ Domain	Engineering	Mathematics	Science
theoretical articles	0	0	0
empirical studies – teaching and learning	0	1	1
empirical studies – measurement	0	3	2
secondary data analysis	0	0	0
qualitative studies	0	0	0

De Bock, Verschaffel, Janssens, Van Dooren, and Claes (2003) investigated the influence of self-constructed graphical representations on 8th grade and 10th grade students' performance on geometrical word problems about area and volume. They administered four different versions of the test to participants at each grade level: with or without an authentic context, and with or without an integrated drawing instruction. The impact of pictorial contexts was examined by asking half of the students to draw a geometrical figure described in the problem before solving it. They found that for the authentic context, students who had to make a drawing scored significantly lower than students from the non-drawing groups, $F(1,302) = 19.52$, $p < 0.01$. Some of their qualitative findings indicated that the drawing contexts conflicted with students' implicit norms, expectations, and beliefs about doing mathematics. Students were also less familiar with the drawing-based instructions, which may have contributed to poorer student performance on questions containing drawing contexts. These findings suggest that questions containing pictorial contexts may result in poorer student performance.

Berends and van Lieshout (2009) examined how different types of illustrations influence the speed and accuracy of performance of both good and poor arithmeticians on arithmetic word problems. Participants were 5th grade students from different schools in the Netherlands. Students were assigned to either the good or poor arithmetician group depending on their performance on a standard mathematics achievement test in Dutch. A repeated measures ANOVA, with type of resources as within-subjects factor (four types: bare, useless, helpful, and essential; see Figure 4) and group as between-subjects factor (poor vs. good arithmeticians), was applied separately for the accuracy and speed of performance. The ANOVA for the accuracy test, the main effect of type of illustration, was significant, $F(3,126) = 14.99$, $p < 0.001$, partial $\eta^2 = 0.26$. The accuracy of student performance decreased when students were forced to look at the "essential" illustration (see Figure 4) to find the essential information to solve the math problem as compared to looking at the "helpful" illustration. The ANOVA for the speed test, the main effect of type of illustration, was also significant, $F(3,126) = 48.62$, $p < 0.001$, partial $\eta^2 = 0.54$. Problems with bare illustrations were solved the fastest by students, while problems with essential illustrations took the longest to solve. Their results demonstrate that illustrations can have a detrimental effect on both accuracy and speed of performance, produced by the presence of irrelevant, redundant information or by increasing working memory load.

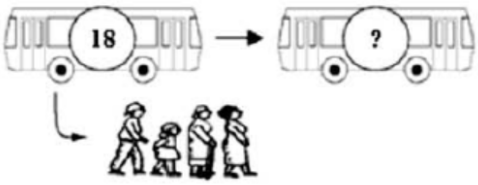
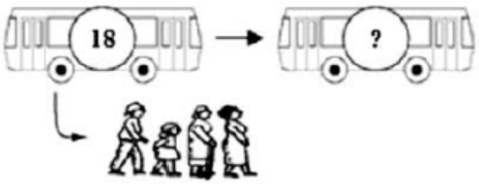
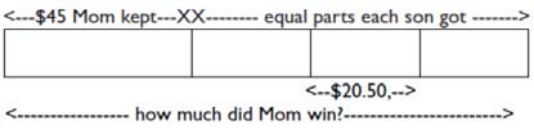
Bare	Useless
There were 18 people on the bus. 4 people got out. How many people are now on the bus? 18 – 4 = ? Answer = …	There were 18 people on the bus. 4 people got out. How many people are now on the bus?  Answer = …
Helpful	**Essential**
There were 18 people on the bus. 4 people got out. How many people are now on the bus?  Answer = …	There were 18 people on the bus. Some people got out. How many people are now on the bus?  Answer = …

Figure 4. Examples of the four types of illustrations from Berends & van Lieshout, 2009, p. 347.

Booth and Koedinger (2012) identified influences of using diagrams and story representations in algebra problems on correct answers by a 3 (grade: 6, 7, 8) × 2 (presentation: story, story along with the diagram) ANOVA design. Participants were students from grades 6-8 at a middle school in the American Midwest. The three sets of items that were the focus of their study were embedded in a written algebra assessment, which had previously been published and used by other research studies (see Figure 5). The ANOVA test indicated a main effect of grade

was found ($F(2, 122) = 4.96$, $p < 0.01$, $\eta^2 = 0.08$), as was a trend toward an interaction between the grade and presentation format ($F(2, 122) = 2.31$, $p = 0.10$, $\eta^2 = 0.04$). Sixth graders answered more story problems correctly than story + diagram problems (i.e., 25% vs. 17%). Seventh and eighth graders tended to answer more story + diagram problem correctly than story problems. Researchers found that while diagrams enhance performance of older and higher ability students, younger and lower-ability student might even be hindered by a diagram's presence.

Story	Story + Diagram
Mom won some money in a lottery. She kept $45 for herself and gave each of her 3 sons an equal portion of the rest. If each son got $20.50, how much did Mom win?	Mom won some money in a lottery, She kept $45 for herself and gave each of her 3 sons an equal portion of the rest. If each son got $20.50, how much did Mom win? (You can use the picture below to help you solve the problem.) 
John bought 3 t-shirts and 2 baseball caps for $58. Sue bought 2 t-shirts and 3 baseball caps for $52. What is the cost of one shirt? What is the cost of one baseball cap?	John bought 3 t-shirts and 2 baseball caps for $58. Sue bought 2 t-shirts and 3 baseball caps for $52. What is the cost of one shirt? What is the cost of one baseball cap? (You can use the picture below to help you solve the problem.) 

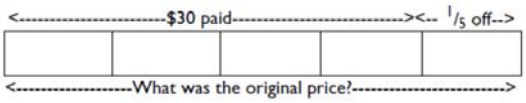
Molly bought a coat on sale. It was 1/5 off the original price. She paid $30. What was the original price of the coat?	Molly bought a coat on sale. It was 1/5 off the original price. She paid $30. What was the original price of the coat? (You can use the picture below to help you solve the problem.) 

Figure 5. Problems displayed in each presentation format from Booth & Koedinger, 2012, p. 497.

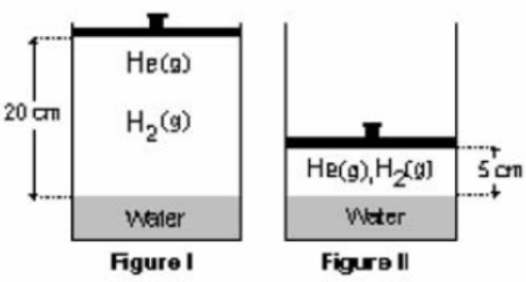
Although prior studies examined these characteristics from different perspectives in mathematics, their findings agreed that pictorial contexts are hard for students. One possible explanation is that students might be unable to extract a correct conceptual understanding of the problem from the visual representations.

In contrast to the findings in mathematics, research findings from science assessments are rather mixed. A study by Costu (2007) examined whether there were significant differences in students' performance on a science assessment when comparing tests with conceptual, algorithmic, or graphical questions (see examples in Figure 6). All three versions of questions included pictorial information. Participants were 11th graders from several different schools of a city in Turkey. Each test included five multiple-choice questions focusing on the topic of gas and gas laws that were published by various item banks. Statistical analysis using one-way ANOVA of student test scores pointed to statistically significant differences amongst each of three test scores ($F(2, 212) = 6.53$, $p < 0.01$) in favor of the conceptual test. In other words, students performed the best on conceptual tests among all types of tests. Also, multiple comparisons suggest that there was a statistically significant difference between the conceptual test and graphical test ($p < 0.001$). However, the three versions of questions differed too much in

their underlying constructs to point to any firm conclusions. For example, the statistical difference might not be solely explained by the inclusion of the graphs in the graphical question. Therefore, the performance difference might be only partially explained by the inclusion of the graphs. Further analyses from students' self-preference reports indicated that most of the students lack graphical understanding and need more training in this area.

Sample of algorithmic problem solving question

Question 1 A:

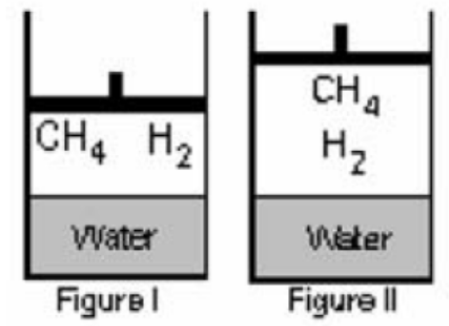


At 27 degree Celsius, total pressure of gasses in the beaker in Figure I is 240 millimeter of mercury (mmHg). What would be the total pressure of the beaker when the piston of beaker is pushed as in Figure II at the same temperature? (Note: Water pressure is 25 mmHg at the 27 degree Celsius)

A) 60 mmHg
B) 360 mmHg
C) 860 mmHg
D) 885 mmHg
E) 960 mmHg

Sample of conceptual understanding question

Question 1 C:

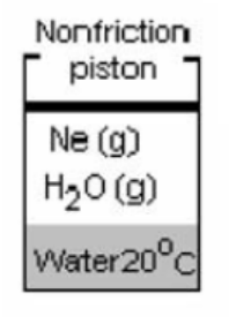


There are gasses of H2, CH4 in a beaker as in Figure I. The piston of beaker is drawn up as in Figure II. After it was reached an equilibrium at the same temperature, how would be changed the pressure of H2, CH4 and water pressure?

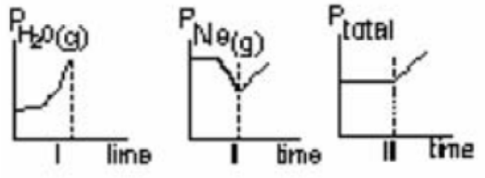
	H2	CH4
A)	Decreases	Decreases
B)	Increases	Increases
C)	Decreases	Decreases
D)	Increases	Increases
E)	Decreases	Decreases

Sample of graphical understanding question

Question 1 G:



There are gasses of Ne and water vapor in a beaker with non-friction piston as in figure. The beaker is gradually heated until 90 degree Celsius. Which of the graph or graphs given below illustrates true changes in this process?



A) Only I
B) Only III
C) I and II
D) II and III
E) I, II and III

Figure 6. Examples of conceptual, algorithmic and graphical questions from Costu, 2007, p. 381.

Turmo and Elstad (2009) analyzed what factors make science test items difficult for majority and minority students. They examined several explanatory variables of test items such as reading load, illustrations, format, writing load, item difficulty, test location, and science subject domain. Participants were 5th and 8th graders in Norway who took the compulsory science standardized achievement test. Majority students were from western countries (e.g., USA,

Canada, Australia, and New Zealand) and minority students were from non-western countries (e.g., Asia, Africa, Latin America, and Oceania). Researchers examined the empirical relationship between Cohen's d and explanatory variables at the item level. They found no relationship between majority and minority student performance gaps and the number of illustrations used in science items.

In contrast, in a recent work on PISA science items, Ruiz-Primo and Li (2012) reported that items with pictorial contexts were easier than those with contexts that included only textual information. Their arguments were consistent with prior research studies about the relationship between textual and non-textual information: (1) environmental society symbols with contextual details result in higher correct comprehension scores from examinees because they reduce ambiguity (Miller, 1995; Wolff, 1996); and (2) built upon cognitive load theory, texts impose greater information processing demands on working memory than pictures or diagrams (Chandler & Sweller, 1991; Sweller, van Merrienboer, & Paas, 1998).

In another science study, Solano-Flores and Wang (2011) carefully defined and presented a conceptual framework for examining pictorial contexts in science assessment. Five main categories were used to describe the presence or absence of different visual features: (1) Representation of objects and background, (2) Metaphorical visual language, (3) Text in illustration, (4) Representation of variables, constants, and functions, and (5) Illustration-text interaction (Solano-Flores & Wang, 2011). They found that illustrations with functional labels provided visual support for ELLs.

In sum, previous research studies in math found that pictorial contexts might hinder the student answering process, thus lowering performance scores. Yet the conclusions on how this characteristic might influence student performance in science are mixed.

2.3.4 Familiar/Unfamiliar Contexts.

In one of four reviewed articles (see Table 4), Chipman, Marshall, and Scott (1991) defined familiarity as the degree of student familiarity with the content in mathematics word problems. Later, Ahmed and Pollitt (2007) similarly defined familiarity as the degree to which the context is familiar to examinees. Familiarity can be specified as test takers' familiarity with settings and culture presented in contexts (Crisp, 2011; Pollitt, Marriott & Ahmed, 2000). In sum, I think familiarity should be defined from the perspective of test-takers rather than the item itself, which consists of three components: settings, phenomena, and culture presented in contexts.

Table 4

Number of Papers Examining Familiarity Characteristics

Type \ Domain	Engineering	Mathematics	Science
theoretical articles	0	0	1
empirical studies – teaching and learning	0	0	0
empirical studies – measurement	0	2	1
secondary data analysis	0	0	0
qualitative studies	0	0	0

In a case study of a student recently immigrated to England from Pakistan who seemed to be underperforming in mathematics, Pollitt, Marriott, and Ahmed (2000) compared his performance on two tests with the data from 39 other students. The boy's right answers to difficult items provided clear evidence that he was indeed quite able at mathematics under some conditions. In order to find out why he got more than half the items on the tests wrong regardless of difficulty,

but correctly answered many difficult items, researchers conducted post-test interviews with that student. They found that his being unfamiliar with the culture and certain scenarios explained why he had difficulty in answering most items he got wrong. For example, the boy reported that he did not know the term "pie chart," and the interviewer inferred that the boy was misled by the phrase "on the packet," in which case the student thought that he was being asked to draw the information on the packet. For another item, the context was situated in posting parcels in British post offices; the boy, who recently immigrated from Pakistan, had almost no experience in post offices and mistakenly thought he should apply a good math technique to interpolate between the charges in the table to find the best cost for the parcel. Therefore, researchers argued that non-familiar contexts could prevent comprehension and task solution, thus resulting in invalid test scores.

However, this feature may not always affects item difficulty in all circumstances because the topic itself, independent of cultures and scenarios, may have an effect. In an empirical study conducted by Crisp (2011), researchers compared student performance on science items between one group of students with reading difficulties and another group of students without reading difficulties. Although study results lack statistical significance, small performance differences between two groups for some items could be linked to the familiarity, via experience, with a context. For instance, an item asked students about the properties of plastic that make it a more appropriate material than iron for making a bucket. Daily life experience with both materials might help both groups of students in answering this question. Setting a question in a familiar context was associated with easiness (Crisp, 2011). The presence of familiar contexts especially

helped students with weaker reading skills by minimizing any language-related barriers and thus allowing them to perform relatively well.

From a theoretical point of view, we need to be aware that any real-world contexts can be more familiar to some students than others. When students are unfamiliar with a particular context involved in an item, they may think the context is to some extent something they have failed to learn, or that the context is something irrelevant to scientific principles (Ahmed & Pollitt, 2007). Considering familiarity based on students' learning experiences, one can more easily hypothesize that some contexts should be more familiar to students than others, as learning experiences might be different across different demographic student groups.

In a recent study examining the instructional sensitivity of contextualized science items as a means of evaluating the instruction, Li, Ruiz-Primo, Wills, and Giamellaro (2012) developed assessment items based on the variations in the proximity of assessments to the enacted curriculum. Researchers considered how familiar the question asked to students was in relation to the instructional activities that students experienced in the science classrooms for a particular module of learning with their science teacher. Interestingly, researchers observed that items with relatively familiar contexts featured as the objects, materials, and procedures similar to what students had experienced in class investigations, yielded a much larger gain of student scores from the pretest to the posttest compared to items with relatively unfamiliar contexts. This finding was consistent across different science modules.

2.3.5 Research Gaps on Context Characteristics

After summarizing common characteristics studied in the literature, I found little connection among those characteristics. Theoretical frameworks across those studies were different from

each other as well. This lack of a consistent theory-driven framework to guide the study of characteristics of item contexts in STEM fields calls for a systematic study that examines how those characteristics are linked with student performance. In addition, most previous studies treated students as a homogenous group and did not look specifically at the ELL subgroup's performance (Wang & Li, 2014). Therefore, in the next section, studies tapping into how item contexts are associated with ELLs' performance are reviewed.

2.4 Relevant Literature on ELL Testing

Putting questions into contexts inevitably involves using extra words (Ahmed & Pollitt, 2007), which may lead to alteration of the item's linguistic demands; such alterations can influence students' performance in ways that are exterior to the intended construct to be measured. Although most of the ELL testing research studies have not explicitly investigated characteristics of contextualized items, several specific research studies focusing on math and science domains are relevant to this study. I reviewed six studies, three from math: Abedi and Lord (2001), Martiniello (2008; 2009), Wolf and Leon (2009); and three from science: Siegel (2007), Ilich (2013), Wang, Liaw, Li, and Taylor (2014).

2.4.1 Math Contextualized Problems for ELLs

In terms of the linguistic demands of test items, regardless of whether the items are contextualized or non-contextualized, previous research indicates that the complexity of texts plays an important role in student test performance (Shaftel et al., 2006). Abedi and Lord (2001) theorized several linguistic features that could be revised in math word problems to improve student performance by reducing linguistic demands. Example features included the following revisions: (1) unfamiliar or infrequently used words were changed into familiar and frequent

non-math vocabulary (e.g., a certain reference file → Mack's company); (2) passive verb forms were changed into active (e.g., if a marble is taken from the bag → if you take a marble from the bag); (3) complex question phrases were changed to simple question words (e.g., which is the best approximation of the number → approximately how many); and (4) abstract or impersonal presentations were made more concrete (e.g., radios sold → radios that Mrs. Jones sold).

Abedi and Lord found that the difference in math performance became statistically significant when the linguistically modified and original items were compared with the revised versions. Further, ELL students' scores on the linguistically modified version were slightly improved from original versions, although the impact was not significant. Additionally, in interviews, students indicated their preference for the modified versions because the vocabulary in the revised items was more familiar to them, and revised items reduced unnecessary comprehension difficulties by decreasing linguistic demands.

Similarly in another math study on ELL testing, Martiniello (2008) described linguistic features of math word problems that posed disproportionate difficulty for ELLs. Her study suggested that linguistic demands were derived from syntactic complexity and lexical complexity, such as the number of clauses, noun phrases, verbs, and verb phrases. Expert ratings of the items' overall syntactic and lexical complexity were used. Through both differential item functioning (DIF) statistics and think-aloud transcripts analyses, Martiniello reported that math items with greater linguistic demands created comprehension difficulties for Spanish-speaking ELLs with respect to syntactic and lexical complexity. Further, Martiniello (2009) examined the impact of non-math linguistic complexity, such as symbolic and visual representations, as a source of DIF in math word problem for ELLs. She concluded that the impact of linguistic

complexity on DIF is attenuated when items provide nonlinguistic schematic representations with which relationships among numbers and variables are represented. This finding suggests that the inclusion of schematic representations in items could help ELLs make meaning of the text and lead to a moderate increase in ELL scores.

In Wolf and Leon's study (2009), they investigated whether the use of academic vocabulary was associated with those DIF items favoring ELLs after controlling for the content difficulty. Academic vocabulary consisted of three subcategories: general academic, context-specific, and technical. General academic vocabulary is defined as words or phrases that can be used across multiple disciplines such as *consequently* or *based on*. Context-specific vocabulary includes that which is used in particular disciplines with specific meanings such as *gas*, *liquid*, and *sound*. Technical vocabulary comprises discipline-specific terminology such as *hypotenuse* or *square root*. Their study yielded a strong association between the use of academic vocabulary and DIF statistics for ELLs, with such vocabulary being a disadvantage in relatively easy items but an advantage in the "not easy" items. To be more specific, when an item requires relatively easy content knowledge, the number of general academic vocabulary terms could disadvantage ELL students, especially the ones with low English language proficiency. In contrast, when an item requires difficult content knowledge ("not easy" item), context-specific and technical vocabulary aided the ELLs.

2.4.2 Science Contextualized Problems for ELLs

Siegel (2007) investigated classroom assessment modifications for ELLs in science courses. Participants were 8th graders including both non-ELLs and advanced ELLs from two diverse middle schools in California. Two versions of assessments were developed and tested with a

pretest/posttest design. Changes to items were made, including adding visual supports (e.g., including a picture of a labeled cough syrup bottle and spoon), and dividing prompts into smaller units (e.g., reducing the number of words). The modified version included not only changes to language, but also visual changes and cognitive changes. Regression analyses of raw and Rasch modeled data suggested that both non-ELLs and advanced ELLs scored significantly better on the modified classroom assessments.

In a recent study of state science tests, Ilich (2013) examined how linguistic complexity as a source led to low performance scores of Spanish-speaking ELL students in grade 5. Features of linguistic complexity included multi-meaning words, academic language, verb phrases, conditional clauses, and so forth. Through an item-level examination of science items that favored non-ELLs, non-familiar words and multi-meaning words were found to have a negative effect on ELL performance.

However, in another study of the Washington Assessment of Student Learning (WASL) science contextualized items, Wang, Liaw, Li, and Taylor (2014) found that the greater the linguistic demands of science item contexts, the more the items favored ELLs. In their study, linguistic demands presented in contextualized items were defined as reading difficulty (e.g., vocabulary and sentence structures) and process difficulty (e.g., ideas). They found that linguistic demands were positively related to ELLs' test scores, possibly because greater linguistic demands of contexts offered ELLs more scientific clues, thus providing ELL students with a better understanding of the scenarios needed to answer the questions.

Findings on ELL testing informed my study design in three ways. First, sequence of contextual information is one unique item context characteristic that is closely linked with text

features such as cohesion, which might have a potential impact on different demographic student groups. Second, previous studies on linguistic features helped me identify what item contexts are more linguistically demanding than others. I could manipulate levels of linguistic demands as high and low in my study, considering text features such as lexical diversity, syntactic complexity, and cohesion. This allowed me to investigate how linguistic demands of contextualized items can differentially affect subgroups. Third, previous studies suggest that I should use comparable constructs in my study, which refer to science items with similar cognitive processing levels. In most secondary data analysis studies, different science items tapped into different levels of cognitive processing skills. Some items might test students' conceptual understanding, some might test students' scientific reasoning skills, and some might test students' science investigation design. Findings resulting from different constructs were incomparable. Thus, it is important to use science items with similar cognitive processing levels to assess student understanding and generate reliable and valid findings.

2.5 Gaps in Research around Item Contexts

Prior research has accumulated some understanding of what characteristics of item contexts are associated with student performance, but those studies are unsystematic and the results are mixed. This study filled four research gaps in the literature around contextualized items.

First, previous studies lack a consistent theory-driven framework for item contexts to guide the characteristics being studied. That may explain why existing studies have no connection to each other, and it may also explain why studies examining similar characteristics resulted in mixed findings. For instance, Berends and van Lieshout (2009) found that illustrations could have a detrimental effect on both accuracy and speed of student performance, produced by their

supplying irrelevant or redundant information or by increasing working memory load. However, Ruiz-Primo and Li (2012) reported that items supplemented with pictorial contexts were easier for students to comprehend than those with contexts as textual information. Inconsistent theoretical frameworks across the two studies led to incomparable findings, resulting in an urgent need for a consistent, theory-driven framework in the STEM assessment field.

Second, prior studies tended to treat students as a homogeneous group. Many of those studies did not specify how context characteristics impact high performers or low performers, or impact ELLs or non-ELLs, respectively. Thus, results across different studies were mixed. To illustrate mixed findings for one particular characteristic – pictorial contexts – De Bock et al. (2003) concluded that graphical representations yielded a negative effect on mainstream students' performance on geometry word problems about area and volume. Yet Solano-Flores and Wang (2011) reported an opposite finding for ELLs. They found that ELLs slightly benefited from illustrations with functional labels, which could have provided visual support for better understanding item contexts. Therefore, it is important to examine subgroups' performance rather than treating students as homogeneous.

Third, most previous studies did not look at how a certain context characteristic such as focusedness interacted with other relevant item features such as item format or linguistic demands (Wang & Li, 2014). It remains unclear whether or not a particular context characteristic is mediated by other item features. My study bridged this research gap by incorporating item features such as levels of linguistic demands.

Finally, prior studies focused on examining various characteristics of contexts without considering the nature of scientific inquiries in the STEM field. Scientific inquiries are primarily

based upon experimental investigations, reasoning, or observations, which are naturally sequential to some degree. But how the sequence of information provided in item contexts influences students' answering process has never been examined empirically.

In order to close those research gaps, my study articulates the theoretical framework for dimensions of context sequence, field tests developed science items with middle and high school students, and conducts student cognitive interviews to examine how patterns of sequential cues in item contexts are linked to student performance of different subgroups (e.g., ESLs, ELLs, non-ELLs).

Chapter III: Methods

In this section, I first present a theoretical framework about the sequence of information in science item contexts. This theoretical framework guided the item selection and development so that the item sequences could be systematically varied for experimental purposes. Then, I describe the research plan, which includes (1) an experimental design in which students are randomly assigned to selected items, and item scores are analyzed to study whether the sequence of contextual information impacts student performance; and (2) cognitive interviews to explore how students perceive the sequence of contextual information presented in the science items. I incorporated the linguistic demands of contexts in my analysis in order to understand whether and how linguistic demands of contexts mediate the relationship between sequence patterns and ELL and non-ELL student performance on contextualized science items.

3.1 Item Development

3.1.1 Operationalizing Dimensions of the Context Sequence

Cognitive psychology research has indicated that a well-substantiated model or definition of the construct improves construct validity, resulting in improved student understanding of what the items are testing (Gorin & Embretson, 2006). The way that students actually reason in the answering process is a very important and relevant issue for researchers in that it helps understand student responses beyond the test results of specific scientific skill applications and knowledge assessments. As noted in the human reasoning theory proposed by Evans et al. (1996), humans do not automatically begin to reason until certain cues activate their thinking process. Since many scientific facts, such as those arrived at via investigations or observations, are presented by using sequential cues, examining what and how sequential cues in item contexts

are linked with student performance differences through understanding reasoning processes may be useful in efforts to improve test items.

Sequence of contextual information is built upon Hoey's (2005) theory of lexical priming. Priming is one of the factors that influences the accessibility of information in memory (Hogg & Cooper, 2003). Providing prime-relevant information can make the activation of stored knowledge more accessible in memory and reduce the cognitive load in problem solving. There are several ways to reduce cognitive load by priming, each of which helped guide the theorization of context sequence in my study.

First, context sequence involves idea ordering and structuring a text (linearizing; Coirier, Favart, & Chanquoy, 2002). Mayer and Moreno (2003) recommended the use of key words to organize information for cognitive processing, and suggested that using key words as lexical priming might facilitate the process of selecting and organizing relevant information in problem solving. Based on previous studies, I hypothesize using key words as logical connectives (e.g., first, second; Tyler, 1994) in item contexts to signal the ordering of ideas might be easier for students in problem solving, thus resulting in better performance in test items.

Second, context sequence involves connectives and textual relationships, and one such relationship is purpose summarization (Graesser, McNamara, & Louwerse, 2003). Bassok and Novick (2012) discussed that when people attempt to find ways to reach their purposes, they draw on a variety of cognitive resources and engage in a host of cognitive activities. Explicitly providing an intention to summarize the topic in item contexts could potentially invites students to consider the relevance of perception and background knowledge to problem solving. Based

on theories in psychology, I hypothesize providing an intention to signal the purpose connectives could reduce the cognitive load in problem solving, thus resulting in better student performance.

Third, another text-connecting relation involved in context sequence is causal relationship (Graesser, McNamara, & Louwerse, 2003). Duncker (1945) noted that identifying the causes of events is one of the goal-directed activities that might facilitate problem solving. Graesser et al. (2003) further argued that using causal mechanisms that explain how mechanisms work, procedures for accomplishing objectives, and logical justifications of claims could offer readers more world knowledge about the ideas in the texts. Built upon previous studies, I hypothesize providing a cause to signal the causal coherence relation could lead to better student performance.

To this end, I theorized three dimensions to capture the context sequence this characteristic: sequence of events, sequence of intention and action, and sequence of cause and effect.

3.1.1.1 Sequence of events (D1)

The first dimension, *sequence of events* (D1), refers to a certain pattern of time or space with which a set of events or objects in contextualized items is presented. Sequence of events can be either *cyclical* or *linear*. Items that involve cyclical events without clearly defined starting and ending points or stages, such as a water cycle, nitrogen cycle, and carbon cycle, are categorized as cyclical sequences. In contrast, item contexts either following or failing to follow a time or space sequence are categorized as linear sequences. This study focuses on linear sequences only.

A linear sequence can be *ordered* (i.e., following either a temporal or a spatial pattern) or *non-ordered* (i.e., NOT communicating either a temporal or a spatial pattern although information involved is temporal or spatial in nature). An ordered sequence of events can signal the descriptions of objects or the order of the presented events, actions, or steps. For example,

linear orderings involve chains of statements that include words such as *first*, *second*, or *later* that give people clues with which to draw inferences (Newstead, Bradon, Handley, Evans, & Dennis, 2002). Theoretically speaking, *sequence of events* (D1) can be decomposed as in the following diagram (see Figure 7). In this study, I am interested in testing the effect of linearly ordered items with and without key words on student performance. I hypothesized that linearly ordered items with key words were easier than those without key words.

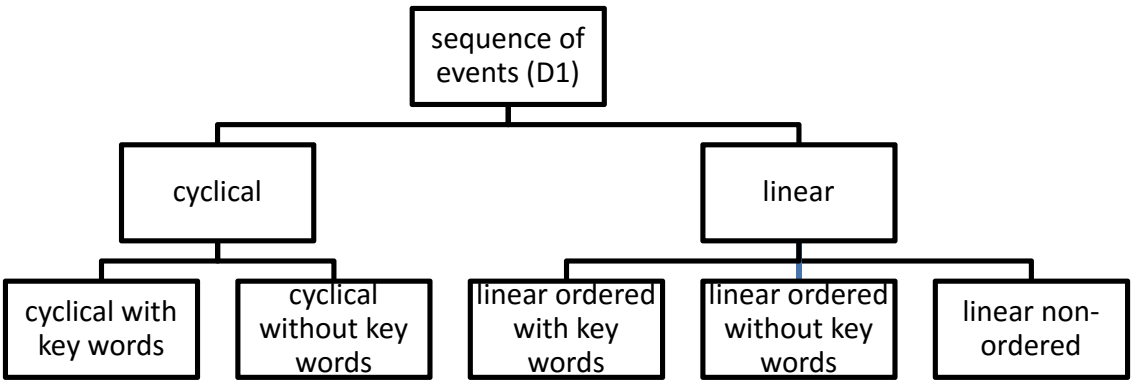


Figure 7. Structure of sequence of events.

Figure 8 shows one pair of examples in the *sequence of events* dimension (D1). To make an ordered item with key words, as shown as version 1 (V1), it is recommended that key words such as *first*, *second*, *later*, *before,* or *after* be added to indicate the orderings of events. Similarly, to make an ordered item without key words, as shown as version 2 (V2), using key words that give hints of the orderings of events should be removed. Contexts of a pair of items in D1 differ only in the presence or absence of key words.

Ordered Item with Key Words (V1)	Ordered Item without Key Words (V2)
❀ **First**, male mussels release sperm into the water. **Then** female mussels take the sperm into their gill chambers where fertilization occurs. **Later**, young mussel larvae are released into the water where they float freely until they attach to the gill of a host fish. **After** a few weeks, young mussel larvae reach the juvenile stage and drop off. The parasitic behavior of the larvae benefits the mussel in two ways. One benefit is that the fish provides nutrition for the larvae when they are attached to its gill. What is the second way this behavior enhances the survival of the mussel species?	❀ Male mussels release sperm into the water. Female mussels take the sperm into their gill chambers where fertilization occurs. Young mussel larvae are released into the water where they float freely until they attach to the gill of a host fish. In a few weeks, they reach the juvenile stage and drop off. The parasitic behavior of the larvae benefits the mussel in two ways. One benefit is that the fish provides nutrition for the larvae when they are attached to its gill. What is the second way this behavior enhances the survival of the mussel species?

Figure 8. Two example items for sequence of events (D1).

Note. Original Item Source came from the Ohio Released State Science Test (see Appendix A).

3.1.1.2 Sequence of intention and action (D2)

The second dimension, *sequence of intention and action* (D2), refers to whether intention and action are presented in a context, and the order in which they are presented (see examples in Figure 9). The intention appears in the form of a scientific hypothesis or purpose, and the action in the form of investigation plans, procedures, or manipulation of experiments or engineering solutions. Intention helps students get a sense of what variables (i.e., manipulating and outcome) should be paid attention to, or on what basis information presented should be used or judged

(Leighton & Gokiert, 2005). I hypothesized that when an intention is provided prior to action, items would be easier for students than when action is described without any intention.

To create an item with both intention and action, as shown in V1 below, adding one sentence to indicate the action's purpose or a scientific hypothesis is recommended. In Figure 9, the sentence *Julio wanted to know how his pulse rate changed when he ran very fast* presents an intention of the action in item contexts. Notice that both the manipulating variables of running speed and pulse rate are pointed out in this sentence. Similarly, to make an item with action only, as shown in V2 below, the sentence indicating the purpose of an action or investigation should be removed. Yet in order to balance the item length, a sentence describing a general purpose for a series of actions needs to be added, such as "*Julio did an investigation about…*", which only mentions the topic of this item. The contexts of a pair of items in D2 differ only in presenting or not presenting an intention.

An item with both intention and action (V1)	An item with action only (V2)
❀ **Julio wanted to know how his pulse rate changed when he ran very fast**. He measured his pulse rate before he started running, while he was running, and two minutes after he stopped running. Which graph best shows how Julio's pulse rate changed?	❀ Julio did an investigation about pulse rate change. He measured his pulse rate before he started running, while he was running, and two minutes after he stopped running. Which graph best shows how Julio's pulse rate changed?

Figure 9. Two example items for sequence of intention and action (D2).

Note. Original Item Source came from Released NAEP Science Assessment (see Appendix A).

3.1.1.3 Sequence of cause and effect (D3)

The third dimension, *sequence of cause and effect* (D3), refers to whether cause and effect are presented in a context, and the order in which they are presented. Causes often appear in the form of an explanatory model, a theory, or an elaboration of a theory. Effects appear in the form of predictions, observations, experimental results, or descriptions of phenomena. I hypothesized that item contexts with an effect that also provide a cause were easier than versions in which only an effect is presented.

To make an item with both effect and cause, as shown in V1 below, it is recommended one sentence be added to indicate the theory of a scientific effect. In Figure 10, the sentence "*Each house is built in a way that helps George and Felice live in their climate*" presents the theory of an inhabitant phenomenon (i.e., how climate affects house designs) in item contexts. Similarly, to make an item with an effect only, as shown in V2 below, the sentence explaining the theory of

the scientific phenomena should be removed. Contexts of a pair of items in D3 differ only in their presenting or not presenting a cause.

The two example items from Figure 10 were piloted in Spring 2014. Of the 396 participating students, 186 took the item on the left (V1), and 210 students took the item on the right (V2). 1-PL item response theory (IRT) analysis reported that the left item with both cause and effect presented (V1) was easier than the right item with only effect presented (V2) in terms of item difficulty (i.e., -0.92 vs. -0.40 on the logit scale, respectively).

An item with both effect and cause (V1)	An item with effect only (V2)
❀ George lives in a house that is located in a very cold, wet climate. Felice lives in a house that is located in a very hot, dry climate. **Each house is built in a way that helps George and Felice live in their climate**. George's house contains a heating system. Felice's house contains an air-conditioning system. Which of the following statement is NOT correct?	❀ George lives in a house that is located in a very cold, wet climate. Felice lives in a house that is located in a very hot, dry climate. Their houses are built in different ways. George's house contains a heating system. Felice's house contains an air-conditioning system. Which of the following statement is NOT correct?

Figure 10. Two example items for sequence of cause and effect (D3).

Note. Original Item Source came from Released NAEP Science Assessment (see Appendix A).

In a preliminary study, I applied this theoretical framework to analyze three dimensions of context sequence with the NAEP science assessment items (Wang, Li, Thummaphan, & Ruiz-Primo, 2013). My colleagues and I found that NAEP contextualized items had at least one sequence dimension described above, and some had more than one sequence dimension. However, items in some versions for a given dimension were less common than others. For example, in D3, item contexts that describe the effect (V2) are more commonly used than items

that describe both effect and cause (V1). In other words, items similar to the example on the right in Figure 10 more frequently appear than items similar to the example on the left.

Further, items used in the NAEP tests (and most likely, in any state test) were not developed based on explicit theories on contexts and even less on the context sequence. Lack of a sufficient sample of items within some dimensions in the context sequence made it difficult to link certain sequential characteristics with student performance. Thus, it is impossible to make recommendations as to which version of context sequence is easier for students, based on the secondary data analysis. This suggested the need for an experimental design to test the effects of the context sequence.

In sum, this theoretical framework benefits item development and guides the research design. To shed light on one benefit, the present study systematically compared differences in student performance by comparing paired items, in which only one dimension of the context was manipulated at a time, keeping all the other characteristics the same.

3.1.2 Item Development

First, I planned the number of science items needed in this study. Then, I carefully selected released science items that could potentially fit the theoretical framework of sequence of contextual information described above, which was developed by my colleagues and me in previous work (Wang, Li, Thummaphan, & Ruiz-Primo, 2013). Later, I followed guidelines to construct two versions of the items for each of three sequence dimensions: sequence of events, sequence of intention and action, and sequence of cause and effect. Last, I discussed those items with content experts and linguists to review, revise, and finalize science items. Detailed steps are described below.

3.1.2.1 Step 1: Number of science items

Within each dimension, I planned to construct two pairs of items with low linguistic demands and another two pairs with high linguistic demands, totaling twelve pairs of items ($n_{item} = 24$). All the items are intended to measure the science content and process skills based on the science curriculum content taught by the Seattle Public Schools district, the district from which participating English Language Learners (ELLs) and non-ELLs were recruited. Item topics include life science (LS) and physical science (PS), both of which are consistent with the district science curriculum and aligned with the Next Generation Science Standards.

Reasons on developing four pairs of science items per dimension are as follows:

(1) Constraints of practical operational work. Power analysis suggests for medium effect size item pairs (roughly 0.50-0.55), I need about 30 pairs of items for each dimension to detect the significant student performance. For smaller effect size, a larger item sample size is needed. Considering the constraint of the operational work, such as the testing length, I cannot test students on 90 pairs of items (i.e., 30 pairs * 3 sequence dimensions = 90 pairs) within two class sessions.

(2) Item sample size from previous research studies. The item sample size in my study was suggested by prior empirical studies that examined how item characteristics impact the item difficulty. For example, Ahmed and Pollitt (2001) used three pairs of science items and one pair of math items to detect mean differences between multiple versions of items. Similarly, Hickendorff (2013) used two pairs of items for each math operation (addition, subtraction, multiplication, and division).

Therefore, as an exploratory study, I chose four pairs of science items per dimension in order to achieve a detectable mean difference between the two versions (see Table 5). After

considering different science topics and different levels of linguistic demands of item contexts for each sequence dimension, this study ended up one pair per cell as shown below.

Table 5

Item Development Plan

Dimension	Linguistic Demands: Low		Linguistic Demands: High	
	Life Sci.	Phys. Sci.	Life Sci.	Phys. Sci.
D1: Sequence of events	1 pair	1 pair	1 pair	1 pair
D2: Sequence of intention and action	1 pair	1 pair	1 pair	1 pair
D3: Sequence of cause and effect	1 pair	1 pair	1 pair	1 pair

Note. Levels of linguistic demands were judged by a panel of experts and double-checked with Coh-Metrix indices.

3.1.2.2 Step 2: Revision of released items

For the test selection process, I selected candidate items from pairs of test questions taken from the released set of middle school NAEP, TIMSS, PISAs or state science tests. Released items have been empirically tested and found to be of sound technical quality, so using released items as sources in this study was more efficient than developing new items. Items with contexts in options were NOT considered in this study because the constructs did not fit in the theoretical framework of the context sequence.

Table 6 provides information indicating the original source of each item (see Appendix A for original items). Item statistics (i.e., p-value) based on a national or international sample are available from online sources such as the NAEP website or TIMSS technical reports. For example, the original source of a physical science item with low linguistic demands in D2: *Sequence of cause and effect* was from a released open-ended NAEP item for Grade 12 in Year 2005. The original item's p-value is 0.14.

Table 6

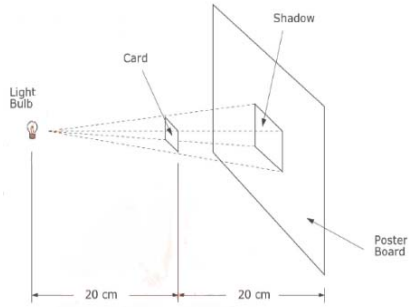
Original Item Sources

Dimension	Linguistic Demands: Low		Linguistic Demands: High	
	Life Sci.	Phys. Sci.	Life Sci.	Phys. Sci.
D1: Sequence of events	Michigan YR2008 Gr8	TIMSS YR2003 Gr8	Ohio YR2007 Gr8	PISA YR2006 Gr8
D2: Sequence of intention and action	Louisiana YR2005 Gr8	NAEP YR2005 Gr12 $p = 0.14$ (OE)	TIMSS YR2011 Gr8 $p = 0.68$ (OE)	TIMSS YR2011 Gr8 $p = 0.39$ (OE)
D3: Sequence of cause and effect	TIMSS YR2011 Gr8 $p = 0.58$	PISA YR2006 Gr8 $p = 0.68$	PISA YR2000 Gr8	PISA YR2006 Gr8

3.1.2.3 Step 3: Follow guidelines to manipulate contexts

After identifying the selected item, if it was a multiple-choice item, its counterpart item was constructed by manipulating only one of the sequence dimensions (i.e., D1, D2, D3) to form the pairs. For example, the original source of PS_D1_L was from TIMSS in YR2003 for Gr8 students (see the first item in Figure 11). I followed the theoretical framework for D1: *sequence of events*, and highlighted the manipulations in bold italic font. I formed item pairs as shown in the second and third item in Figure 11. Revisions not relating to the theoretical framework for D1 made by me or panel reviewers were in bold underline font.

A tiny light bulb was held 20 centimeters to the left of a square card, which was in turn held 20 centimeters to the left of a poster board, as shown. The shadow of the card on the poster board had a side of 10 centimeters.

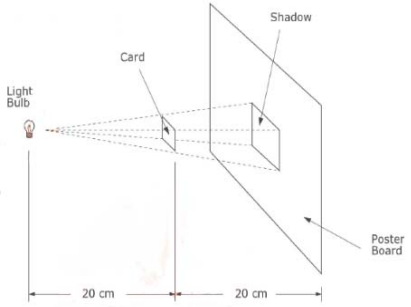


If the poster board is moved 40cm further to the right so that it is 80cm from the light, what will be the new side of the card's shadow on the poster board?

A. 5cm
B. 10cm
C. 15cm
D. 20cm

Released item from TIMSS

Lily conducted an investigation on the light. A tiny light bulb was held 20 centimeters to the left of a square card, which was in turn held 20 centimeters to the left of a poster board, as shown. The shadow of the card on the poster board had a side of 10 centimeters.



She moved the poster board 40cm further to the right. So the poster board was 80cm from the light and 60cm from the card.

What will be the new side of the card's shadow on the poster board?

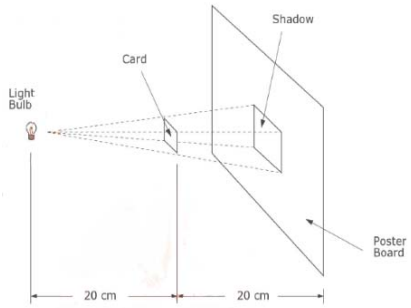
A. 20cm
B. 15cm
C. 10cm
D. 5cm

Add a description sentence to follow the operationalized guidelines for D1: Sequence events.

Break one sentence into two to make it clear.

Re-order options to balance the keys in the whole tests.

Lily conducted an investigation on the light. ***First***, a tiny light bulb was held 20 centimeters to the left of a square card, which was in turn held 20 centimeters to the left of a poster board, as shown. The shadow of the card on the poster board had a side of 10 centimeters.



Second, she moved the poster board 40cm further to the right. So the poster board was 80cm from the light and 60cm from the card.

What will be the new side of the card's shadow on the poster board?

A. 20cm
B. 15cm
C. 10cm
D. 5cm

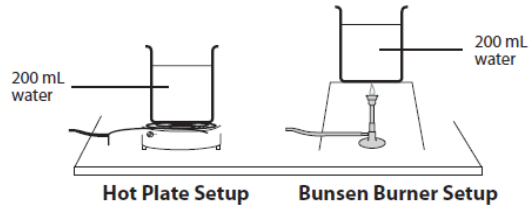
Form its counter version in the test.

Follow the operationalized guideline in D1: sequence of events, to add key words indicating the sequence of steps.

Figure 11. Examples of forming a multiple-choice item pair in D1: sequence of events.

Due to the test length limit and resource constraints in scoring, I constructed only multiple-choice items in this study. For selected open-ended items, I first converted the item into a multiple-choice format and then created its counterpart item. For example, the original source of PS_D2_H was from TIMSS assessments YR2011 for Gr8 students (see the first item in Figure 12). I changed the prompt from a statement into a question and formed four options based on TIMSS technical reports (see the second item in Figure 12). Last, I formed item pairs shown in the second and third item based on the theoretical framework for D2: *sequence of intention and action,* and highlighted the manipulations in bold strikethrough font. Revisions not relating to the theoretical framework for D2 made by me or panel reviewers were in bold underline font.

Two kinds of heat sources are usually available in the science lab; an electric hot plate and a Bunsen burner. Jack planned an investigation to test which of these sources heats water faster. He poured 200 mL of water into each of two identical beakers and recorded the initial temperature of the water in each beaker. Jack then placed one beaker on a hot plate and the other over a Bunsen burner, as shown below.



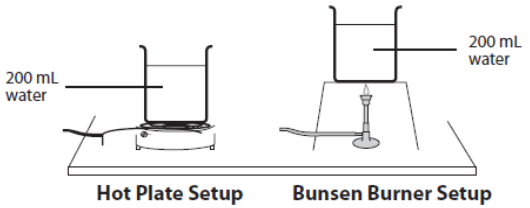
He recorded the temperature of the water in each set up every two minutes for ten minutes.

List one variable that Jack controlled in his investigation.

Scoring rubric:
Correct Response
• Lists one variable as shown below.
The beakers (same, same shape, same size, same materials)
The water (same volume, from the same place)
The thermometer (same type, same position for taking readings)
Location of the experiment (same place, same room)
Incorrect Response
• Incorrect (including crossed out, erased, stray marks, illegible, or off task)
Examples:
The initial temperature.
Checking the temperature.
Timing.

Released item from TIMSS

Two kinds of heat sources are usually available in the science lab; an electric hot plate and a Bunsen burner. Jack planned an investigation to test which of these sources heats water faster. He poured 200 mL of water into each of two identical beakers and recorded the initial temperature of the water in each beaker. Jack then placed one beaker on a hot plate and the other over a Bunsen burner, as shown below. **He recorded the temperature of the water in each set up every two minutes for ten minutes.**



What variables did Jack control in his investigation?

A. Beakers and the initial temperature of the water.
B. Timing and the temperature of heat sources.
C. Beakers and the amount of water.
D. The temperature of heat sources and the location of the experiment.

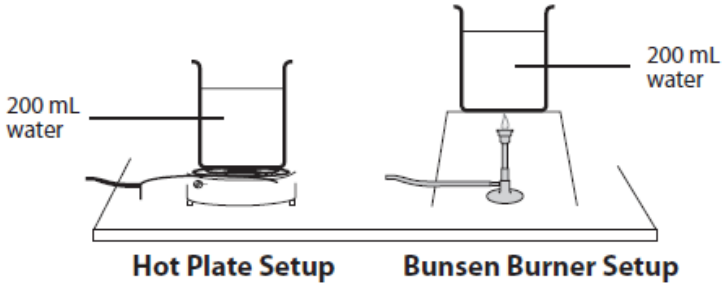
Item ID: PS_D2_V2_L1

Move the position of this sentence to make the format flow better.

Change the question to make the difficulty level aligned with the NGSS.

Add options based on the examples of correct responses provided by TIMSS.

Two kinds of heat sources are usually available in the science lab; an electric hot plate and a Bunsen burner. ~~Jack planned an investigation to test which of these sources heats water faster.~~ Jack poured 200 mL of water into each of two identical beakers and recorded the initial temperature of the water in each beaker. Jack then placed one beaker on a hot plate and the other over a Bunsen burner, as shown below. He recorded the temperature of the water in each set up every two minutes for ten minutes.



What variables did Jack control in his investigation?

A. Beakers and the initial temperature of the water.
B. Timing and the temperature of heat sources.
C. Beakers and the amount of water.
D. The temperature of heat sources and the location of the experiment.

Item ID: PS_D2_V1_L1

Form its counter version in the test.

Follow the operationalized guidelines in D2: sequence of intention and action, to remove a sentence indicating the purpose of the investigation.

Figure 12. Examples of forming a multiple-choice item from an open-ended item in D2: Sequence of intention and action.

Guided by operationalized definitions for each dimension described previously, I was able to specify item-writing rules during this item development process.

3.1.2.4 Step 4: Final review

I collaboratively worked with two experienced science content experts (one from life science and one from physical science) with solid content knowledge to review and revise six pairs of items for each dimension; that is, two items in any pair differing only in a particular dimension.

After the initial items were constructed, as described in Step 3, I conducted a content panel review. The panel review included the two science content consultants and one measurement professor to ensure that items were scientifically accurate and used grade-level appropriate language. Science items were modified according to the panel's recommendations. For example, one revision (break one sentence into two to make it clear) for the second item in Figure 11 was suggested in the panel review.

Levels of linguistic demand within each dimension were judged by a panel of experts (i.e., one linguist, one Gr8 science teacher, and one bilingual English/Spanish teacher) in Fall 2014. The professional panel compared two pairs of science items in the same topic from the same manipulated dimension (e.g., two pairs of LF items from D1), and then classified one pair into linguistic low level and the other pair into linguistic high level. Experts' classification was based on three indices (i.e., lexical density, sentence complexity, and discourse complexity).

Afterward, I checked whether the levels of linguistic demands (high vs. low) aligned with experts' judgments, as reported by Coh-Metrix Common Core Text Ease and Readability Assessor (T.E.R.A.) (Crossley, Greenfield, & McNamara, 2008). Coh-Metrix T.E.R.A. was used only as an analytic tool for me to better understand linguists' classification criteria.

Coh-Metrix T.E.R.A. is designed to analyze the easability and readability of texts and provides useful information about text features. It analyzes texts on five components: narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. Based on the T.E.R.A website, the descriptions of the five components are describe as follows: (1) Narrativity: The more story-like a text the higher the narrativity score, and the easier the text. (2) Syntactic simplicity: Texts with fewer clauses, fewer words per sentence, and fewer words before the main verb will have a higher score for syntactic simplicity. (3) Word concreteness: A text with relatively high numbers of concrete words is easier to read and will have a high word concreteness score. (4) Referential cohesion: Texts with overlap between words, word stems, or concepts from one sentence to another have referential cohesion; use of more similar words or conceptual ideas makes it easier for readers to make connections between ideas. (5) Deep cohesion: A measurement of how well the events, ideas, and information of the whole text are tied together.

I used two life science items from D1: *Sequence of events* (see Figure 13) to illustrate how Coh-Metrix T.E.R.A provided the analysis on the linguistic demands of item contexts. I interpreted the results and classified one item into linguistic low group and the other one into linguistic high group. First, I input item contexts of two items into the Coh-Metrix T.E.R.A online system (http://www.cohmetrix.com/).

LS_D1_01 Kim wanted to determine if certain seeds require sunlight to germinate. First, she placed one seed in a moist paper towel in the sunlight. Then, she placed another seed in an equally moistened paper towel in a dark closet. After a few days, the seed in the sunlight germinated but the one in the closet did not. Last, Kim reported to the class that this type of seed needs sunlight in order to germinate.
LS_D1_02 First, male mussels release sperm into the water. Then female mussels take the sperm into their gill chambers where fertilization occurs. Later, young mussel larvae are released into the water where they float freely until they attach to the gill of a host fish. After a few weeks, young mussel larvae reach the juvenile stage and drop off. The parasitic behavior of the larvae benefits the mussel in two ways. One benefit is that the fish provides nutrition for the larvae when they are attached to its gill.

Figure 13. Two life science items from D1: sequence of events.

Then, I ran the text analysis and comparison for those two item contexts (see Figure 14). Obviously, LS_D1_01 is easier than LS_D1_02 in terms of the five text components. For instance, LS_D1_01 is much easier than LS_D1_02 (i.e., 65% vs. 20%) in terms of narrativity. Also, the ideas presented in LS_D1_01 are a little more closely and smoothly tied than are the ideas presented in LS_D1_02 (i.e., 78% vs. 68%). Details of the automated analysis are reported in Figure 14.

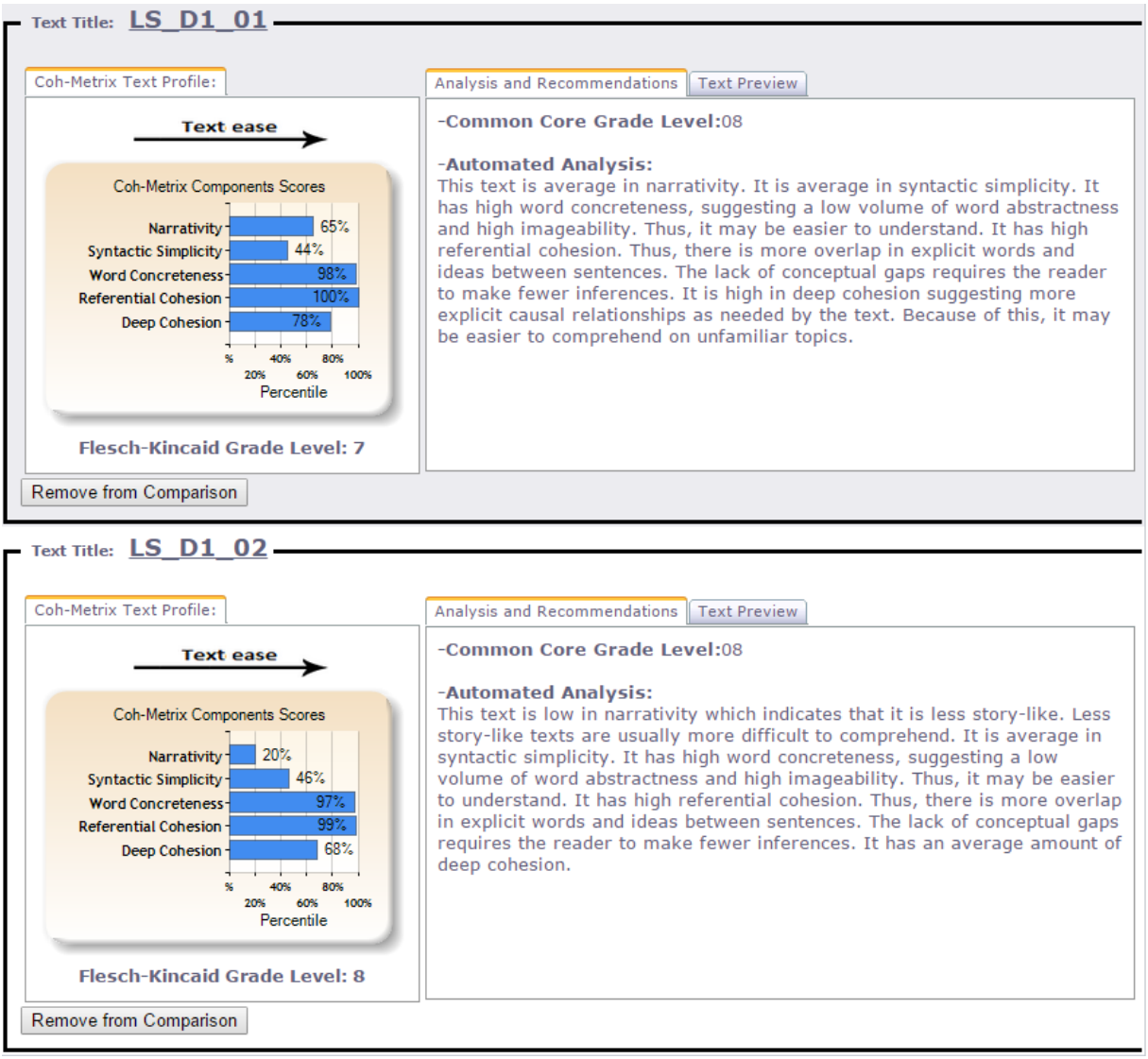


Figure 14. T.E.R.A. text comparison.

3.2 Participants

In Spring 2015, finalized items were field tested with a sample of 974 middle school and high school students from China and the U.S. to investigate how the various dimensions of context sequence impact student performance. Of the participating students, 617 were English as Second Language Learners (ESLs) from China, 103 were English Language Leaners (ELLs) from the U.S., and 254 were non-ELLs from the U.S. This number of participants allows for a

strong base for analysis from a psychometric perspective, ensuring that there was enough data on each science item (Wright & Stone, 1979). Further, I used a set of items to conduct cognitive interviews with a stratified sample of 18 students in the U.S and 9 students in China. The purpose of these cognitive interviews was to examine students' cognitive processes that occur while they are responding to items.

3.2.1 ESL participants from China

According to ETS online resources, ESL refers to students who study English as a second or foreign language, live in non-English-speaking countries, and are in an instructional or school setting where English as a foreign language is not used. In this study, most ESLs started to learn English at school from pre-K or Grade 1. Therefore, the average length of time ESL students have been learning English at school is about 9-10 years. Throughout the study, I refer to the Chinese students as ESLs.

Participants from Beijing. Two schools in Beijing participated in this study: Beijing Haidian Foreign Language Experimental School and Beijing RDFZ Chaoyang School. Both schools are first-class schools in Beijing and use small classroom size as a teaching strategy, with the number of students per class at around 30. Participants included 269 students from 10 classes at Grade 9 in Haidian and 145 students from 5 classes at Grade 10 in Chaoyang.

Participants from Guangzhou. One school in Guangzhou participated in this study: The High School attached to the Sun-Yatsan University. It is a first-class high school in Guangdong province. The school places great emphasis on science and technology education, with characteristic programs such as scientific and technological invention experiments, robotic

experiments, projects using computers, and hands-on projects and inventions. Participants included 203 students from 4 classes at Grade 10.

3.2.2 U.S. Teacher Background

Five teachers – three females and two males – from four school districts in the Greater Seattle Area voluntarily signed up for this study. Among the four school districts, two teachers came from a suburban district and three came from urban districts. Participating schools all expressed a great desire to improve students' science learning. Table 7 provides the teachers' demographic information (e.g., experience, education, gender).

Table 7

Teacher Demographics

Characteristics	Teachers	
School district location	Suburban	Urban
Number of teachers	2	3
Gender	1 Female, 1 Male	2 Females, 1 Male
Education	2 MA	2 MA, 1 BA
Average years of teaching	4 (Range: 1-7)	16 (Range: 10-20)
Other information	2 science teachers from one middle school	1 science teacher from one middle school 1 science teacher from one high school 1 ELL teacher from one high school

3.2.3 ELL and non-ELL Participants from the United States

In total, 357 students at Grades 7 through 12 from five teachers' classrooms in the Greater Seattle Area participated in this study. Students and teachers were recruited from the Seattle Public Schools district, which has an active working relationship with the College of Education

at the University of Washington. At the school district level, there are 43.2% White, 19.0% Asian and Pacific Islanders, 12.3% Latino, 18.6% African American, 1.2% Native American, and 5.7% Multiracial students, with 40.5% qualifying for free/reduced lunch and 12.5% eligible for ELL services. Recruited participants were from two middle schools and two high schools. Of the 357 students, 103 were ELLs and 254 were non-ELLs. According to the federal government, ELL refers to students with limited English proficiency who are acquiring English as part of their education [public law 107-110, Title IX, Part A, Sec 9101, (25)]. Throughout the study, I refer to the U.S. students as ELLs and non-ELLs.

For research purposes, ELLs in my study were oversampled – most ELLs were recruited directly from ELL classes in two high schools. The 103 ELLs comprised a mixture of White, Asian, Latino, African American, and Multiracial students. Student demographic information was collected from the participating teachers. This information includes student gender, race, and English Language Learner designation.

3.3 Data Collection

3.3.1 Procedure for Field Tests

I used the two-booklet design to randomly assign each pair of items into one of two booklets (i.e., 12 items per booklet; see booklet design map in Table 8). In the table, item ID: LS_D1_V1_LinguisticLow means life science item from Dimension 1 version 1 with low linguistic demands. Each pair of two items was randomly assigned into one of the two test booklets. Using this approach, only one version of a pair of items appears in each booklet. Several factors were considered when ordering the items: (1) Items with low linguistic demands were used as questions 1-6, and items with high linguistic demands were used as questions 7-12;

(2) items that covered different science topics (i.e., life science and physical science) were, ideally, arranged one after another; (3) items were fit onto one page for the pencil-paper version of the booklets; and (4) items with pictorial resources such as diagrams, tables, or pictures were arranged after the items without pictorial resources.

Table 8

Booklet Design Map

Item No.	# of Students*	Booklet 1 (BK1)	Booklet 2 (BK2)
1	974	LS_D1_V1_LinguisticLow	LS_D1_V2_LinguisticLow
2	974	PS_D2_V2_LinguisticLow	PS_D2_V1_LinguisticLow
3	974	LS_D2_V2_LinguisticLow	LS_D2_V1_LinguisticLow
4	974	PS_D3_V1_LinguisticLow	PS_D3_V2_LinguisticLow
5	974	LS_D3_V2_LinguisticLow	LS_D3_V1_LinguisticLow
6	974	PS_D1_V1_LinguisticLow	PS_D1_V2_LinguisticLow
7	974	PS_D3_V1_LinguisticHigh	PS_D3_V2_LinguisticHigh
8	974	LS_D1_V1_LinguisticHigh	LS_D1_V2_LinguisticHigh
9	974	PS_D2_V2_LinguisticHigh	PS_D2_V1_LinguisticHigh
10	974	LS_D2_V2_LinguisticHigh	LS_D2_V1_LinguisticHigh
11	974	PS_D1_V1_LinguisticHigh	PS_D1_V2_LinguisticHigh
12	974	LS_D3_V2_LinguisticHigh	LS_D3_V1_LinguisticHigh

**Including both Chinese and American students.*

Note. Item ID was given by Item Topic Dimension No._Version No._LinguisticLevel.

All students were administered the two booklets in two field tests, one booklet first and another booklet after three to four weeks. Further, I included two testing orders of the booklets to control for the possible ordering effect or learning effect of the test (see Figure 15). That is,

half of the students were randomized to initially use the first booklet, and the other half to initially use the second booklet; after three to four weeks, those students who had already taken the first booklet took the second booklet, and vice versa.

This crossover administration method (Piantadosi, 2005) combined with booklet designs allowed me to: (1) have fewer students than non-crossover designs; (2) have students take all the items without noticing the obvious similarity or difference between two items in each pair, thus reducing the possible response bias effect; and (3) balance the task and make the test booklet a reasonable length to avoid fatigue.

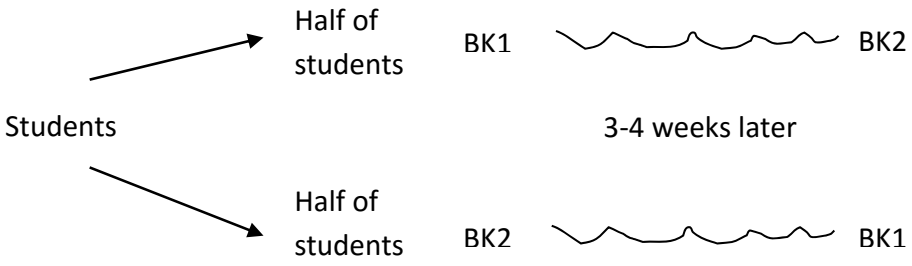


Figure 15. Field test design.

The three- to four-week test break was suggested by the curve of forgetting (Finkenbinder, 1913). The curve accounts for how we retain or get rid of information that we take in, based on a one-hour lecture. For example, by Day 7, we will have forgotten 90% of what we learned. By Day 20, we retain about 2%-4% of what we learned in the original one-hour lecture. This nicely coincides with the first field test, and may account for students reporting that they felt as if they had never seen the items before when they were taking the second field test. Students have to actually re-do the test from scratch.

In practice, some classes or some students could not participate in the second field test. I summarized the number of students for every booklet in Figure 16. In total, there were 699 students who finished BK1, and 808 students who finished BK2. Among all 974 students, there were 533 students who finished both booklets. Based on the information in Table 9, about half of the students were male and the other half female for both booklets. Students and teachers answered several some survey questions for this study (see details in Appendix D and Appendix E).

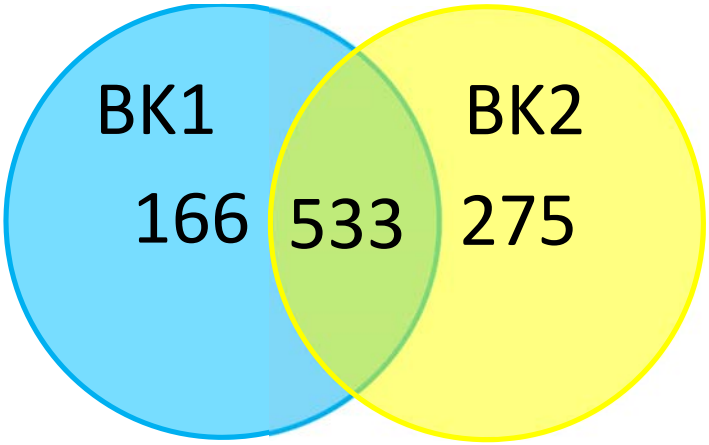


Figure 166. Participants for two booklets.

Table 9

Gender Frequency Table

Booklet / Gender	BK1		BK2	
	Frequency	Percent	Frequency	Percent
Boys	354	50.64	408	50.50
Girls	338	48.35	396	49.01

3.3.2 Procedure for Cognitive Interviews

A stratified sampling strategy was used to identify U.S. students and Chinese students from Beijing for the cognitive interview. Prior to selecting the students for the interviews, I grouped the U.S. students and Chinese students from Beijing into high, medium, and low performers based on their first field test performance on the test items. Next, a random stratified sample of 27 students was selected for the cognitive interviews (see Table 10). I took into consideration the balance of student gender and performance level for each pair of items. I randomly selected 3 non-ELLs, 3 ELLs, and 3 ESLs, per performance stratum of test performance level (high-, middle-, and low-performance). I conducted 27 student cognitive interviews, using both concurrent and retrospective approaches, with the 18 students (i.e., 9 ELLs and 9 non-ELLs) from the Greater Seattle Area and 9 ESLs from Beijing, China to collect more direct evidence on students' cognitive processes when responding to items.

Table 10

A Stratified Sample of Students

Performance level / Dimension	High	Medium	Low
D1: Sequence of events	1 non-ELL; 1 ELL; 1 ESL	1 non-ELL; 1 ELL; 1 ESL	1 non-ELL; 1 ELL; 1 ESL
D2: Intention and action	1 non-ELL; 1 ELL; 1 ESL	1 non-ELL; 1 ELL; 1 ESL	1 non-ELL; 1 ELL; 1 ESL
D3: Effect and cause	1 non-ELL; 1 ELL; 1 ESL	1 non-ELL; 1 ELL; 1 ESL	1 non-ELL; 1 ELL; 1 ESL

Students were interviewed regarding their thoughts about all pairs of items, that is, 4 pairs of items for each dimension with a total of 12 pairs. Specifically, within each dimension, the

selection was taken into account regarding whether the items showed desirable or undesirable patterns of the sequence dimension and linguistic demand on student performance. Due to the length of the interview, 12 pairs of items were organized into three sets so that each student was randomly assigned to one set, with one to two pairs of items showing the hypothesized pattern in student performance while the other pairs showed the opposite effect (see item pair assignment details in Appendix C). Therefore, each pair of items had interview responses from 9 students (3 high, 3 medium, and 3 low performers). For each dimension, four pairs of items had 36 interview responses (9 students * 4 pairs = 36 student responses).

The sample size of students and the length of the interview were informed by other studies. For example, Ahmed and Pollitt (2010) interviewed 11 students with two think-aloud type questions. Leighton and Gokiert (2005) conducted an hour interview to invite students to think aloud concurrently as they solved the items. Students in my study were selected for interviews by me based on their first field test performance score or were recommended by their teachers, based on their classroom performance.

The cognitive interview (i.e., the concurrent and retrospective interviews) was carried out individually after the second administration of the test booklet. Each student was given a test book of four pairs of items and asked to respond to a series of questions about each item (see below for the example questions). Students were videotaped to capture the audio while also videotaping the test page that students were responding to. Interviewers also took notes during the interview. Both videotaped interviews and observational notes were used as sources of data about students' cognitive processes.

The interviews were conducted with protocols that had been used successfully in interview studies on contextualized items (Ahmed & Pollitt, 2007; Almond, et al., 2009; Ericsson & Simon, 1993; Ferrara et al., 2004; Hamilton et al., 1997; Karabenick & Woolley, 2006; Leighton & Gokiert, 2005; Noble et al., 2012). In the interviews, students were asked to first read the item aloud, then verbalize the process that they used to approach the item, and finally respond to several questions asking about strategies they employed or confusion they experienced. The interview protocols for the present study are as follows:

(1) Read aloud, "*We are trying to find out what students of your age think about science questions that appear on tests. In a moment, I am going to be showing you some science questions. As you read each one, I want you to tell me what is going through your mind.*" (Ericsson & Simon, 1993; Leighton & Gokiert, 2005).

(2) Verbalization of the process, "*Also, as you answer each science question, I want you to tell me what is going through your mind. Do you think you can do this?* (Practice with students by asking them to think aloud on one math item: 19*5). *Please try very hard to talk about what you are thinking. If I notice that you have stopped talking, I will remind you to keep talking.*" (Ericsson & Simon, 1993; Leighton & Gokiert, 2005). To remind students to keep talking, we use the prompt "*Please remember to say everything that is going through your mind*" (Ericsson & Simon, 1993; Leighton & Gokiert, 2005) or "*What are you thinking about right now? Can you tell me more?*" (Almond et al., 2009).

(3) After a student finishes verbalizing her/his problem-solving process, the interviewer continues asking follow-up questions to clarify how the student interpreted and answered the questions. The following are some example questions: "*How were these two items similar or*

different? (Ahmed & Pollitt, 2007); "*Which item in this pair is easier for you to understand, and why?*"; "*Which part helps/confuses you when answering the question?*"; "*What could we do to the item to make the question clearer for you or your peers?*"

3.4 Data Coding and Analyses

The quality of items were first examined under the Classical Test Theory (CTT) approach, including *p* values, corrected item-total correlations, and Cronbach's alpha if an item was deleted. The p-value from the CTT approach is the proportion of examinees responding in the correct direction, equivalent to the mean of item scores (Allen & Yen, 2002). Corrected item-total correlation measures the relationship of individual items to the overall test score excluding the targeted item. A low correlation means the item is not measuring the same thing the rest of the test is trying to measure. Cronbach's alpha coefficient is used as an estimate of the reliability of a psychometric test. By using this definition, it is implicitly assumed that the average correlation of a set of items is an accurate estimate of all items that pertain to a certain construct (Allen & Yen, 2002).

3.4.1 Data Analysis for RQ1

To answer the **first research question**, *How is providing the sequence of information presented in item contexts associated with student performance?*, I first performed an IRT model selection (1-PL, 2-PL, 3-PL) for dichotomous items for all participants, and then performed three individual IRT runs to compare item parameters within item pairs for each dimension for ESLs (n = 617), ELLs (n = 103), and non-ELLs (n = 254). Last, for three different demographic groups, I performed multiple t-tests for 12 pairs of item difficulty estimates and standard errors of the estimates to test whether there is a significant difference between pairs.

IRT is a statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test (Hambleton & Russell, 1993). There are three dichotomous IRT models: 1-parameter logistic model (1PL), 2-parameter logistic model (2PL), and 3-parameter logistic model (3PL). For example, in a 3PL model, the probability of a correct response is: $p(\theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}}$, with b as the location on the underlying scale, where examinees have a 50/50 chance of responding correctly; a as the discrimination; and c as the pseudo chance or guessing that a low ability examinee will get the correct answer (Hambleton, Swaminathan, & Rogers, 1991). A 2PL model assumes that both difficulty b and discrimination a can differ, and there is no guessing parameter c. A 1PL model assumes that only difficulty b differs, and that the discrimination parameter a is the same for all examinees regardless of their location on the underlying trait, and there is no guessing parameter c.

The purpose of model selection is to choose a model that not only provides sound fit to the data but also has the ability to generalize to predictions of future data (Kang & Cohen, 2007). To that end, statisticians have proposed information theoretic measures to evaluate model fit. One popular measure is the Bayesian information criterion (BIC), which provides estimates of the relative differences between solutions (Schwarz, 1978). The BIC criterion is defined as:

$$BIC(M) = D(M) + \log(n) \times \left| M \right|$$

where the first component, $D(M)$, is the deviance, and it will always decrease as more parameters are added to the model; n is the sample size; and $\left| M \right|$ is the number of estimated parameters, and it penalizes for the complexity of a model. In other words, the second

component will always increase as you add more parameters to the model. The first and second components together create a balance between fit and complexity. Overall, BIC gives a higher penalty to the number of parameters and tends to choose models with fewer parameters. Thus, smaller values of BIC are preferred among models. I used the **ltm** package in R to perform the model selection and IRT analysis.

After the model selection, I performed three individual IRT runs on three sub-data sets: one for ESLs, one for ELLs, and a third one for non-ELLs. Then, I compared the parameter *b* within item pairs for three different demographic groups – ESLs, ELLs, and non-ELLs – in two ways: descriptive comparisons and t-tests. The descriptive method directly compared the value of parameter *b*. Twelve t-tests for one sub-group at one time. When fitting an IRT model, it reports the item difficulty estimate (i.e., parameter *b*) and standard errors. In order to test whether the difference between two difficulty estimates is significant, I checked whether the two confidence intervals +/-2.65*SE overlap. I performed twelve t-tests on the same dataset; after the Bonferroni corrections, the cumulative probability *p* should be adjusted to 0.004 (i.e., 0.05/12 = 0.004). Thus, the absolute z-value of 2.65 was applied. Last, I interpreted the patterns and results for each sequence dimension across the three subgroups.

3.4.2 Data Analysis for RQ2

To answer the **second research question**, C*onsidering different levels of linguistic demands in contexts, how is contextual sequence linked to science performance of ELLs and non-ELLs? ESLs and non-ELLs?*, I performed two differential item functioning (DIF) runs for two datasets: ELLs vs. non-ELLs, ESLs vs. non-ELLs. I calculated the DIF estimates for each pair and

qualitatively examined the characteristics of DIF items for two datasets. I used **difR** package in R to perform the DIF analyses.

DIF occurs when test takers from different demographic groups have different probabilities or likelihoods of success on a test item given the same ability. It provides an indication of unexpected behavior of items on a test, for example, favoring one group and disfavoring another group. Mantel-Haenszel (MH) is a common procedure for assessing DIF. The MH procedure is a chi-squared contingency table method for estimating and testing a common two-factor association parameter in a 2 * 2 * K table, where K serves as the basis for matching members of both the reference and focal groups on an individual item (Holland & Thayer, 1988). The contingency table rows refer to either reference or focal groups, and the columns refer to correct or incorrect responses.

The next step in calculating the MH statistic is to use data from the contingency table to obtain an odds ratio, an estimate of $\hat{\alpha}_{MH} = \frac{\sum A_k D_k / N_k}{\sum B_k C_k / N_k}$, for all values of k and where N_k represents the total sample size at the kth interval. The scale of α is 0 to ∞, with α =1 indicating no DIF. The statistical significance of α_{MH} can be tested with a χ^2 test with 1 degree of freedom as $\chi^2_{MH} = \frac{(|\sum_k N_k - \sum_k E(N_k)| - \frac{1}{2})^2}{\sum_k Var(N_k)}$. Holland and Thayer (1985) proposed to use $\Delta_{MH} = -2.35 In(\hat{\alpha}_{MH})$ as a measure of DIF in the scale of differences in item difficulty as measured in the ETS "delta scale." A negative delta scale favors a reference group such as non-ELLs, while a positive delta scale favors focal groups such as ESLs or ELLs.

Zieky (1993) described three categories of DIF magnitude for classifying DIF items. Both χ^2_{MH} (with p =.05) and the absolute value of Δ_{MH} are used for this classification. If the absolute value of Δ_{MH} for a particular item is at least 1.5 and the DIF test is statistically significant (p < .05), the item is classified as having moderate to large values of DIF. If $1.0 \leq |\Delta_{MH}| < 1.5$ and the DIF test is statistically significant, the item is classified as having slight to moderate values of DIF. If the $|\Delta_{MH}| < 1.0$ or the DIF test is not statistically significant, the item is classified as having negligible or nonsignificant DIF (see Table 11).

Table 11.

DIF Classification System

Category	Criteria
Moderate to Large DIF for the focal group	$\Delta_{MH} \leq -1.5$ & $p < .05$
Slight to Moderate DIF for the focal group	$-1.5 < \Delta_{MH} \leq -1.0$ & $p < .05$
Negligible or Nonsignificant DIF	$\lvert\Delta_{MH}\rvert < 1$ or $p > .05$
Slight to Moderate DIF for the reference group	$1.0 \leq \Delta_{MH} < 1.5$ & $p < .05$
Moderate to Large DIF for the reference group	$\Delta_{MH} \geq 1.5$ & $p < .05$

In this study, I performed two DIF runs for two sub-datasets. One sub-dataset contains responses from ELLs (n = 103) and non-ELLs (n = 254). The other sub-dataset contains responses from ESLs (n = 617) and non-ELLs (n = 254). Flagged DIF items for two different runs are reported in the next chapter.

3.4.3 Data Analysis for RQ3

To answer the **third research question**, *How do different dimensions of sequence of context influence how students perceive and respond to tasks?*, I created a codebook and analyzed the transcribed students' interviews supplemented with the interviewer's observation notes.

The codebook used for analyzing the transcriptions contained nine questions (see Table 12). Codebooks were constructed for each of the three different sequence patterns. Every pair of items in each sequence pattern was given to nine students (3 ELLs, 3 non-ELLs and 3 ESLs composed of 3 high performers, 3 moderate performers and 3 low performers). To be more specific, each pair of items was given to both male and female students.

Table 12.

Codebook

ID	Coding Questions
1	What is student background? (e.g., ELL_F_H indicates ELL female high performer)
2	Which version appeared first during the interview? (V1/V2)
3	Did the student select the same option or a different option for the pair? (S/D)
4	Did the student select the correct option for V1? (Y/N)
5	Did the student select the correct option for V2? (Y/N)
6	Did the student notice the difference between the pair?
7	Which pattern did the student comment on in the pair? (V1 is easier than V2, V1 is harder than V2, V1 is similar to V2)
8	What were the student's comments on the pair?
9	What can be done to revise the item?

By coding the transcriptions and quantifying through counting the codes, I analyzed students' responses to concurrent and retrospective interview questions to identify emerging themes that

help to connect the sequence patterns to the statistical analysis and understand the influences of those sequence patterns. These themes were included in my coding categories as shown above. Using both the planned coding categories and emerging codes enabled me to glean useful insights into the reading, reasoning, and thinking processes that students exhibited. I then qualitatively examined reasons underlying the context sequence effects that I observed in the patterns from the results from two previous research questions.

Chapter IV: Results

In this chapter, the quality of items are examined and item statistics are reported. Then, I report findings of the study, ordered by research questions.

4.1 Item Statistics

Item statistics were examined, including *p* values, corrected item-total correlations, and Cronbach's alpha if an item was deleted (see Table 13). The p-value is also called the item difficulty. The range of item difficulties in this study is from .28 to .69, medium difficult to difficult.

Regarding corrected item-total correlations, most items in this study fell into the acceptable range of corrected item-total correlations from .20 to .32. Some items did not fall into this acceptable range, which suggests those items were not measuring the same construct measured by other items included. In other words, test items in this study covered various aspects of science.

Number of students selecting each distractor for each test item is also reported (Table 13), as this data should be considered an important part of the item. The quality of distractors influences student performance on an exam item. Although the correct answer must be truly correct, it is just as important that the distractors be incorrect. Distractors should appeal to students who have not mastered the material, whereas high performers should infrequently select the distractors. Reviewing the distractors could reveal potential errors of judgment and inadequate performance of those distractors. Based on the information reported in Table 13, each distractor was selected by at least a few students. Table 13 also shows the response to total correlation for each distractor. Good distractors should have response to total correlations that are negative or zero.

Also, the average measure of distractors (i.e., the average ability of the people who responded to that distractor) was smaller than the correct options, which suggests that distractors are not appealing to higher performers. Therefore, the quality of distractors was good.

Cronbach's alpha for the 24 items in the present study is .66. Very easy and very difficult items could have low variances, which might affect the alpha coefficient in this study. Cronbach's alpha was calculated by including only students who completed all items. Of those who completed all items, ESLs comprised about 50% (263 out of 533), ELLs comprised about 15% (80 out of 533), and non-ELLs comprised about 35% (190 out of 533).

Table 13

Item Statistics for 24 Science Items

Item No.	Item ID	Difficulty p_i	Original Released Difficulty	A	B	C	D	Corrected Item-Total Correlation	Alpha if Item Deleted
BK1									
1	LS_D1_V1_L	.56	NA	**388** (.16)	52 (-.67)	54 (-.65)	205 (-.32)	.30	.64
2	PS_D2_V2_L	.49	.14 (OE)	179 (-.16)	**343** (.08)	132 (-.32)	45 (-.71)	.08	.66
3	LS_D2_V2_L	.63	NA	27 (-.90)	59 (-.68)	176 (-.40)	**437** (.14)	.32	.64
4	PS_D3_V1_L	.60	.68 (MC)	122 (-.29)	**418** (.12)	95 (-.51)	60 (-.68)	.26	.64
5	LS_D3_V2_L	.43	.58 (MC)	123 (-.56)	95 (-.30)	177 (-.23)	**300** (.22)	.23	.65
6	PS_D1_V1_L	.56	NA	**389** (.12)	125 (-.37)	126 (-.38)	54 (-.47)	.18	.65
7	PS_D3_V1_H	.30	NA	**210** (.25)	72 (-.58)	232 (-.25)	182 (-.13)	.18	.65
8	LS_D1_V1_H	.34	NA	105 (-.28)	**238** (.21)	177 (-.30)	172 (-.24)	.16	.65
9	PS_D2_V2_H	.42	.39 (OE)	116 (-.39)	144 (-.32)	**293** (.27)	138 (-.43)	.28	.64
10	LS_D2_V2_H	.51	.68 (OE)	103 (-.46)	120 (-.43)	117 (-.54)	**351** (.26)	.34	.63
11	PS_D1_V1_H	.33	NA	103 (-.55)	98 (-.19)	260 (-.20)	**229** (.24)	.16	.65
12	LS_D3_V2_H	.43	NA	76 (-.49)	211 (-.22)	**299** (.20)	105 (-.45)	.22	.65

Table 13 (Continued)

Item Statistics for 24 Science Items

Item No.	Item ID	Difficulty p_i	Original Released Difficulty	A	B	C	D	Corrected Item-Total Correlation	Alpha if Item Deleted
BK2									
1	LS_D1_V2_L	.62	NA	**497** (.37)	220 (-.90)	42 (-.80)	49 (-.35)	.31	.64
2	PS_D2_V1_L	.55	.14 (OE)	200 (-.28)	**442** (.34)	128 (-.26)	37 (-.77)	.09	.66
3	LS_D2_V1_L	.69	NA	36 (-.84)	42 (-.95)	169 (-.36)	**559** (.29)	.26	.64
4	PS_D3_V2_L	.66	.68 (MC)	127 (-.30)	**529** (.28)	76 (-.45)	74 (-.64)	.22	.65
5	LS_D3_V1_L	.39	.58 (MC)	134 (-.50)	144 (-.36)	210 (-.01)	**317** (.49)	.30	.64
6	PS_D1_V2_L	.69	NA	**556** (.26)	82 (-.32)	112 (-.55)	53 (-.49)	.22	.65
7	PS_D3_V2_H	.40	NA	**326** (.38)	59 (-.70)	178 (-.15)	245 (-.10)	.13	.66
8	LS_D1_V2_H	.28	NA	113 (-.29)	**229** (.42)	209 (-.07)	253 (-.06)	.19	.65
9	PS_D2_V1_H	.46	.39 (OE)	149 (-.32)	144 (-.37)	**371** (.48)	142 (-.33)	.30	.64
10	LS_D2_V1_H	.55	.68 (OE)	127 (-.29)	134 (-.26)	100 (-.49)	**443** (.35)	.14	.66
11	PS_D1_V2_H	.41	NA	83 (-.41)	129 (-.33)	265 (-.20)	**328** (.49)	.22	.65
12	LS_D3_V1_H	.47	NA	77 (-.48)	256 (-.19)	**374** (.42)	93 (-.41)	.17	.65

Note. [a] Item key is marked in bold red.
[b] Average measure is reported in the parentheses under options.

4.2 Sequence of Contextual Information

For the first research question, *How is providing the sequence of information presented in item contexts associated with student performance?*, I theorized three dimensions in the context sequence. It was hypothesized that adding sequential key words, the intention of the investigation, and the cause of an effect would help students perform better.

I began statistical analysis by first checking the local dependence between pairs of item in Winsteps, then performing the IRT model selection for dichotomous items using **ltm** package in R, and finally, comparing item parameters within item pairs for each dimension for ESLs, ELLs, and non-ELLs.

Since data were collected from the same person more than once and were collected on similar science item pairs, I needed to check the local dependence. In practical terms, a correlation of r = 0.20 is low dependency in that the two items only have 0.2*0.2 = 0.04 of their variance in common (Linacre & Wright, 2000). In this study, 96% of each of their residual variances differed. Linacre and Wright (2000) suggest correlations need to be at least 0.70 or above before we are concerned about dependency. The Winsteps results produced standardized residual correlations to identify dependent items (reported in Table 14; the largest correlations are shown in the table), indicating no dependency between items, and the assumption of IRT was not violated. In this table, we are usually only interested in correlations that approach 1.0 or -1.0, because that may indicate that the pairs of items are duplicative or are dominated by a shared factor (Linacre & Wright, 2000). High positive residual correlations may indicate local item dependency between pairs of items.

Further, I used a factor-analytic approach to detect dimensionality, as suggested in Linacre's paper (2009). One factor size had an eigenvalue of above 2.0. Linacre suggested that a

substantive secondary dimension within the data requires at least two items to define it; thus, a minimum eigenvalue of 2.0 is more meaningful. This finding suggests test items in my study are uni-dimensional.

Table 14

Largest Standardized Residual Correlations

Item Number in BK 1/2	Item	Item Number in BK 1/2	Item	Standardized Residual Correlation
2-BK1	PS_D2_V2_L	2-BK2	PS_D2_V1_L	0.20
4-BK1	PS_D3_V1_L	4-BK2	PS_D3_V2_L	0.16
5-BK1	LS_D3_V2_L	5-BK2	LS_D3_V1_L	0.15
1-BK1	LS_D1_V1_L	1-BK2	LS_D1_V2_L	0.15
2-BK2	PS_D2_V1_L	8-BK2	LS_D1_V2_H	-0.20
2-BK1	PS_D2_V2_L	4-BK1	PS_D3_V1_L	-0.19
2-BK2	PS_D2_V1_L	5-BK2	LS_D3_V1_L	-0.19
1-BK1	LS_D1_V1_L	10-BK1	LS_D2_V2_H	-0.16
2-BK1	PS_D2_V2_L	5-BK1	LS_D3_V2_L	-0.16
2-BK2	PS_D2_V1_L	4-BK2	PS_D3_V2_L	-0.14

Based on the BIC criteria, the 1PL model was preferred as a good fit for the data in contrast to 2PL and 3PL: the smallest BIC was for the 1PL model (see Table 15). As BIC tends to select the simpler model, this finding is consistent with previous work (Lin & Dayton, 1997).

Table 15

Comparisons of Model Selection

Model	BIC
1PL	23514.02
2PL	23580.64
3PL	23680.20

After identifying the 1PL model as the best fitting model for the data, three individual 1PL IRT runs were performed on three sub-data sets: one for ESLs (n = 617), one for ELLs (n = 103), and one for non-ELLs (n = 254). Item parameter *b* for 12 pairs was calculated for ESLs, ELLs,

and non-ELLs respectively. To test whether there was a significant difference between the item parameter *b* between pairs, twelve t-tests were conducted for three sub-data sets individually.

Through the descriptive comparison (comparing the value of item parameter *b*), six pairs showed the hypothesized pattern for ESLs, three pairs showed the hypothesized pattern for ELLs, and two pairs showed the hypothesized pattern for non-ELLs (see Tables 16-18 in red font). Through statistical comparison, three pairs for ESLs showed the significant pattern from D1 to D3 (see Tables 16-18, pairs with star "*"), and no pair showed any significant pattern for either ELLs or non-ELLs.

For D1: *sequence of events*, I hypothesized that adding key words to indicate the order of an investigation (V1) would be easier for students than item contexts without key words (V2). Based on the IRT results reported in Table 16 for three sub-group students, for one physical item, examining the relation between the distance from the light and the size of the shadow, adding key words in item contexts could be significantly detrimental for ESLs.

Also I found within two pairs of life science items, the pair from the linguistic demands: high group were harder for all three sub-groups than the pair from the linguistic demands: low group. Similar results were replicated for two pairs of physical science items.

Table 16

Item Parameter b Comparison for D1: Sequence of Events

Student Status	Linguistic Demands: Low		Linguistic Demands: High	
	Life Sci.	Phys. Sci.	Life Sci.	Phys. Sci.
ESLs (n=617)	V1 > V2 (- .92) (-1.64)	**V1 > V2*** (- .27) (-1.66)	V1 < V2 (1.24) (2.08)	V1 > V2 (1.45) (.67)
ELLs (n=103)	V1 < V2 (1.55) (1.63)	V1 > V2 (- .14) (- .88)	V1 > V2 (1.85) (1.73)	V1 < V2 (1.28) (2.03)
non-ELLs (n=254)	V1 < V2 (- .25) (- .12)	V1 > V2 (- 1.01) (-1.11)	V1 < V2 (.50) (.70)	V1 > V2 (.61) (.16)

Note. * significant difference between pairs (see details in Appendix B).

For D2: *sequence of intention and action*, I hypothesized that item contexts with both intention and action (V1) would be easier for students than item contexts with action only (V2). Based on the IRT results reported in Table 17 for three student sub-groups, adding one sentence to explain the intention of an investigation might help ESLs perform better. One pair of life science items examining structure, function, and information processing showed a significant hypothesized pattern for ESLs. However, those findings were not replicated for ELLs or non-ELLs.

Further, I found within two pairs of life science items, the pair from the linguistic demands: high group were harder for all three sub-groups than the pair from the linguistic demands: low group. Similar results were replicated for two pairs of physical science items.

Table 17

Item Parameter b Comparison for D2: Sequence of Intention and Action

Student Status	Linguistic Demands: Low		Linguistic Demands: High	
	Life Sci.	Phys. Sci.	Life Sci.	Phys. Sci.
ESLs (n=617)	**V1 < V2*** (-2.44) (-1.20)	V1 < V2 (-1.76) (-1.05)	V1 < V2 (- .34) (.22)	V1 < V2 (.37) (.88)
ELLs (n=103)	V1 > V2 (1.17) (.04)	V1 > V2 (1.08) (.96)	V1 > V2 (.66) (.18)	V1 > V2 (1.73) (1.59)
non-ELLs (n=254)	V1 > V2 (- .65) (- .84)	V1 > V2 (- .77) (- .85)	V1 > V2 (- .14) (- .27)	V1 > V2 (- .41) (- .64)

Note. * significant difference between pairs (see details in Appendix B).

For D3: *sequence of cause and effect*, I hypothesized that item contexts with both cause and effect (V1) would be easier for students than item contexts with effect only (V2). Based on the IRT results in Table 18, one pair of physical science items examining states of matters, showed a significant opposite hypothesized pattern for ESLs. Adding one sentence to explain the cause of a science phenomenon appears to have no impacts on non-ELL's performance at all, barely helped ELLs, but could be significantly detrimental for ESLs.

Moreover, I found that within the two pairs of life science items, the pair from the linguistic demands: high group were harder for U.S. students (i.e., ELLs and non-ELLs) than the pair from the linguistic demands: low group, except for ESLs. Yet for two pairs of physical science items, the pair from the linguistic demands: high group were harder for all three sub-groups than the pair from the linguistic demands: low group.

Table 18

Item Parameter b Comparison for D3: Sequence of Cause and Effect

Student Status	Linguistic Demands: Low		Linguistic Demands: High	
	Life Sci.	Phys. Sci.	Life Sci.	Phys. Sci.
ESLs (n=617)	V1 > V2 (1.25) (.97)	V1 > V2 (- .27) (- .73)	V1 < V2 (- .21) (.43)	**V1 > V2*** (1.59) (.54)
ELLs (n=103)	V1 < V2 (.91) (1.15)	V1 > V2 (.08) (- .80)	V1 > V2 (1.54) (.83)	V1 > V2 (1.60) (1.44)
non-ELLs (n=254)	V1 > V2 (- .28) (- .82)	V1 > V2 (-2.23) (-2.24)	V1 > V2 (.74) (.23)	V1 > V2 (.99) (.73)

Note. * significant difference between pairs (see details in Appendix B).

4.3 DIF Results

For the second research question, *Considering different levels of linguistic demands in contexts, how is the contextual sequence linked to science performance of ELLs and non-ELLs? ESLs and non-ELLs?*, I ran two DIF analyses. In the first DIF run for ELLs (n = 103) and non-ELLs (n = 254), there were 7 flagged DIF items favoring non-ELLs (see Table 19). Negative delta scale favors non-ELLs, while positive delta scale favors ELLs. Slight to moderate DIF items were noted as effect size B, and moderate to large DIF items were noted as effect size C.

Table 19

DIF Results by Size of DIF Statistics for ELLs and non-ELLs

Item	Stat.	P-value	deltaMH	Effect Size	Favoring
LS_D1_V1_L*	4.12	0.04	-1.02	B	NON-ELL
LS_D1_V2_L*	6.12	0.01	-1.35	B	NON-ELL
PS_D1_V2_H	6.12	0.01	-1.40	B	NON-ELL
LS_D2_V1_L	4.94	0.03	-1.16	B	NON-ELL
PS_D2_V1_H*	7.76	0.01	-1.60	C	NON-ELL
PS_D2_V2_H*	6.72	0.01	-1.24	B	NON-ELL
PS_D3_V1_L	7.53	0.01	-1.30	B	NON-ELL

Note. * indicates this item is from a pair.

For D1: *sequence of events*, three items were flagged as DIF items favoring non-ELLs. Among the three DIF items, two items were from a pair. This pair (LS_D1_V1_L, LS_D1_V2_L) examined how to improve an experiment to strengthen the conclusion that certain seeds require sunlight to germinate. Another DIF item favoring non-ELLs was identified in the V2 version (item contexts without key words). This item (PS_D1_V2_H) examined what substances are soluble in water. The manipulated pair were both flagged as DIF items, which may suggest that D1: *sequence of events* tends not to be the source of DIF.

In D2: *sequence of intention and action*, three items were flagged as DIF items favoring non-ELLs. Among the three DIF items, two items were from a pair. This pair (PS_D2_V1_H, PS_D2_V2_H) examined student understanding of controlled and uncontrolled variables in an experiment to test which source heats water faster. Another DIF item favoring non-ELLs was in the V1 version (item contexts with both intention and action). This item (LS_D2_V1_L)

examined student understanding of how to analyze, interpret data, and draw a conclusion from an experiment by adding a certain chemical to water. Similarly to D1, both items in the manipulated pair were flagged as DIF, which may indicate that the source of DIF tends not to be D2: *sequence of intention and action*.

In D3: *sequence of cause and effect*, there was one item in the V1 version (i.e., item contexts with both cause and effect) flagged as a DIF item favoring non-ELLs. This item (PS_D3_V1_L) examined student understanding of how temperature changes the state of water. Since no pairs of items were flagged as DIF items in this sequence dimension, I suspect that D3: *sequence of cause and effect* could be a potential source of DIF.

For the second DIF run (cultural DIF), there were 6 items favoring ESLs and 1 item favoring non-ELLs (see Table 20). The negative delta scale favors non-ELLs, while the positive delta scale favors ESLs. Slight to moderate DIF items were noted as effect size B, moderate to large DIF items were noted as effect size C. It is interesting to find that most DIF items favoring ESLs were moderate to large DIF items.

Table 20

DIF Results by Size of DIF Statistics for ESLs and non-ELLs

Item	Stat.	p-value	deltaMH	Effect Size	Favoring
LS_D1_V1_L*	11.68	0.0006	1.01	B	ESL
LS_D1_V2_L*	38.45	0	2.30	C	ESL
LS_D2_V1_L	36.74	0	2.25	C	ESL
PS_D2_V1_L*	117.93	0	4.13	C	ESL
PS_D2_V2_L*	48.84	0	2.15	C	ESL
PS_D3_V1_L	5.41	0.02	-1.42	B	NON-ELL
LS_D3_V1_H	18.00	0	1.50	C	ESL

Note. * indicates this item is from a pair.

For D1: *sequence of events*, two items from a pair were flagged as DIF items favoring ESLs. This pair (LS_D1_V1_L, LS_D1_V2_L) examined how to improve an experiment to strengthen the conclusion that certain seeds require sunlight to germinate. As with results of the first and second DIF runs, both items in the manipulated pair were flagged as DIF, suggesting that the source of DIF tends not to be D1: *sequence of events*.

For D2: *sequence of intention and action*, three items were flagged as DIF items favoring ESLs. Among three DIF items, two items were from a pair. This pair (PS_D2_V1_L, PS_D2_V2_L) examined student understanding of the relation of boiling temperature to air pressure or altitude. This pair had four options with high linguistic demands in terms of syntactic simplicity compared to the other items. Another DIF item favoring ESLs was in the V1 version (i.e., item contexts with both intention and action). This item (LS_D2_V1_L) examined student understanding of how to analyze, interpret data, and draw a conclusion from an

experiment by adding a certain chemical to water. Again, results suggest that D2: *sequence of intention and action* tends not to be the source of DIF due to both items in the manipulated pair being flagged as DIF.

For D3: *sequence of cause and effect,* there was one item in the V1 version (i.e., item contexts with both cause and effect) flagged as a DIF item favoring non-ELLs and another DIF item in the V1 version favoring ESLs. The flagged DIF item favoring non-ELLs (PS_D3_V1_L) examined student understanding of how temperature changes the state of water. The flagged DIF item favoring ESLs (LS_D3_V1_H) examined student understanding of the food web in ecosystems. Both flagged DIF items in this sequence dimension may suggest that D3: *sequence of cause and effect* can be a potential source of DIF, although it is unclear about the direction of DIF.

A comparison of results for two DIF runs (Table 19 and Table 20) yielded four common science items that were flagged as DIF items (see Table 21). The two DIF runs flagged three life science items as DIF items favoring non-ELLs and ESLs. One pair (LS_D1_V1_L, LS_D1_V2_L) examined how to improve an experiment to strengthen the conclusion that certain seeds require sunlight to germinate. Another item (LS_D2_V1_L) examined student understanding of how to analyze, interpret data, and draw a conclusion from an experiment by adding a certain chemical to water.

One item (PS_D3_V1_L) was flagged as a DIF item favoring the reference group (i.e., non-ELLs) in two DIF runs. This physical science item examined student understanding of how temperature changes the state of water.

Table 21

Comparison of Two DIF Runs

Item	First DIF Run (ELLs vs. non-ELLs)		Second DIF Run (ESLs vs. non-ELLs)	
	Effect Size	Favoring	Effect Size	Favoring
LS_D1_V1_L*	B	NON-ELL	B	ESL
LS_D1_V2_L*	B	NON-ELL	C	ESL
LS_D2_V1_L	B	NON-ELL	C	ESL
PS_D3_V1_L	B	NON-ELL	B	NON-ELL

4.4 Student Cognitive Interviews

To answer the third research question, *How do different dimensions of sequence of context influence how students perceive and respond to tasks?*, I analyzed student interviews to examine which version from each pair of items the interviewed students preferred (i.e., which they found to be easier to understand) and the reasons for their preferences, sorted by sequence dimensions.

4.4.1 D1: Sequence of Events

For each pair of science items in D1, I interviewed nine students (3 ELLs, 3 non-ELLs, and 3 ESLs). There are four pairs of science items in D1; thus, I collected 36 student responses in total. About 11% (4 out of 36) of student responses indicate that they could not distinguish the difference between the pair until they were encouraged by the interviewer to look more closely at the questions. About 67% (24 out of 36) of interviewed student responses mentioned that V1 (i.e., item contexts with key words) was easier than V2 (i.e., item contexts without key words). Due to the limited number of interviewed students, there is no significant difference among

students within the three comparison groups, that is, among ESLs, ELLs, and non-ELLs, between males and females, and among high/medium/low performers, in terms of preferences (i.e., the version of the question that students identified in the interview as easiest) for the pair in D1 (see Table 24).

During the cognitive interview, most student responses (29 out of 34) were the same option for the pair. If students chose the correct option for V1, they would tend to choose the same correct option for V2. Later, I tracked 36 student responses on test items. Only 16 had the same option for the pair, while 20 did not have the same option. In other words, those students' responses in the interview were inconsistent with their responses in the pencil-paper tests. For example, in the interview, most students chose the same option (e.g., A) for the pair. Yet in their pencil-paper tests, only 9 student responses chose A for the pair, while the rest chose different options for the pair (e.g., A for V1, D for V2.).

The landscape of student interviews for D1 is provided in Table 22.

Table 22

Landscape of Student Interviews for D1: Sequence of Events

Item	Item Difficulty Preference								
	V1 < V2 (n=24)					V1 = V2 (n=6)	V1 > V2 (n=4)		Inaudible (n=2)
	Reason 1: V1 connected ideas to each other.	Reason 2: It was easier to follow the steps of the investigation in V1.	Reason 3: V1 signaled the steps/stages of an investigation/phenomenon.	Reason 4: V1 helped students better understand the timing of the events.	NA [a]	Reason 5: V1 and V2 were similar.	Reason 6: V2 had fewer words and went right straight to the point.	NA	
LS_D1_L		1 ELL 1 non-ELL		2 ELLs 1 non-ELL 2 ESLs	1 non-ELL	1 ESL			
LS_D1_H	1 ELL	1 ESL	3 non-ELLs			1 ESL		1ESL	2 ELLs
PS_D1_L		1 ELL 1 non-ELL 1 ESL			2 ELLs 1 non-ELL	1 non-ELL 1 ESL	1ESL		
PS_D1_H	1 ELL		1 ESL		1 ELL 1 non-ELL 1 ESL	1 non-ELL 1 ESL	1 non-ELL	1 ELL	
All items by ELLs	2	2	0	2	3	0	0	1	2
All items by non-ELLs	0	2	3	1	3	2	1	0	0
All items by ESLs	0	2	1	2	1	4	1	1	0

Note. [a] NA indicates the cases when the comments given were not related to the manipulation of the sequence dimensions.

I observed that interviewed non-ELLs preferred V1 (i.e., item contexts with key words) mainly for two reasons: (1) key words made it easier to follow the steps of the investigation and (2) key words signaled the steps of an investigation. Interviewed ELLs preferred V1 because of three reasons: (1) key words connected ideas, (2) key words made it easier to follow the steps of the investigation, and (3) key words helped students understand the timing of events. For example, one ELL student said that "it helps people to sort out all the steps." Another ELL student said "transition words helped me know the stages and the order." Interviewed ESLs preferred V1 mainly for two reasons: (1) key words made it easier to follow the steps of the investigation, and (2) key words helped students understand the timing of events. In addition, I found that the increase of linguistic demands in item contexts did not influence students' perceptions on the item difficulty preference in D1: *sequence of events*.

4.4.2 D2: Sequence of Intention and Action

For each pair of science items in D2, I also interviewed nine students (3 ELLs, 3 non-ELLs, and 3 ESLs). There are four pairs of science items in D2; thus, I collected 36 student responses in total. About 11% (4 out of 36) of student responses indicate that the interviewees could not tell the difference between the pair before they were prompted to more closely investigate the pair. About 56% (20 out of 36) of students indicated that V1 (i.e., item contexts with both intention and action) was easier than V2 (i.e., item contexts with action only). There is not much difference in preferences for the pair in D2 between ELLs and non-ELLs, between males and females, and among high/medium/low performers.

During the cognitive interviews, most interviewed student responses (35 out of 36) selected the same option for both versions during the interview. If students selected the correct option for

V1, they selected the same correct option for V2. Later, I tracked those 36 students' responses on their test items; 21 had the same options for the pair, while the remaining 15 did not have the same options for the pair.

The landscape of student interviews for D2 is provided in Table 23.

Table 23

Landscape of Student Interviews for D2: Sequence of Intention and Action

Item	Item Difficulty Preference							
	V1 < V2 (n=20)			V1 = V2 (n=5)		V1 > V2 (n=10)		Inaudible (n=1)
	Reason 1: V1 helped focus on core components to solve the problem.	Reason 2: V1 gave information what the investigator wanted to do.	NA [a]	Reason 3: V1 and V2 were similar.	NA	Reason 4: V2 was shorter and easier to understand.	NA	
LS_D2_L	1 ELL 2 ESLs	1 ELL 2 non-ELLs			1 ELL	1 non-ELL	1 ESL	
LS_D2_H	1 non-ELL 1 ESL	1 ELL 1 ESL			1 ESL	1 ELL 1 non-ELL	1 non-ELL	1 ELL
PS_D2_L			3 ELLs 2 non-ELLs			1 non-ELL	3 ESLs	
PS_D2_H	1 non-ELL	1 non-ELL 2 ESLs	1 non-ELL	2 ELLs	1 ESL	1 ELL		
All items by ELLs	1	2	3	2	1	2	0	1
All items by non-ELLs	2	3	3	0	0	3	1	0
All items by ESLs	3	3	0	0	2	0	4	0

Note. [a] NA indicates the cases when the comments given were not related to the manipulation of the sequence dimensions.

I observed ELLs, non-ELLs and ESLs who preferred V1 (i.e., item contexts with both intention and action) over V2. They explained that additional manipulation of the sentence indicating the intention of an investigation in V1 did the following: (1) helped them focus on core components to solve the problem, and (2) gave them more ideas of what was happening and what the investigator planned to do. For example, one student mentioned "The manipulation sentence tells me its core variables, and states what Andrea [the investigator] is looking for. So it helps me focus on those components of it instead of photosynthesis as a whole. Since she's telling us about light intensity and carbon dioxide, I know what to look for." Another student explained that "V1 is easier because it said what Jack [the investigator] was trying to do. It gives me more of an idea of what's happening during the experiment and what I should keep controlled and what shouldn't."

Interestingly for physical science items, non-ELLs all changed their item difficulty preferences from preferring V2 to preferring V1 (i.e., they indicated that V1 is easier than V2) when the linguistic demands in item contexts increased. Therefore, the increase of linguistic demands in item contexts for physical science items, in the form of adding the manipulation sentence in V1, appears to influence non-ELLs' perceptions on the item difficulty preference in D2: *sequence of intention and action.*

4.4.3 D3: Sequence of Cause and Effect

For each pair of science items in D3, I also interviewed nine students (3 ELLs, 3 non-ELLs, and 3 ESLs). There are four pairs of science items in D3; thus, I collected 36 student responses in total. About 11% (4 out of 36) of student responses indicate that students could not tell the difference between the pair before the interviewer encouraged the students to look harder at the

questions. About 64% (23 out of 36) of student responses indicated that V1 (i.e., item contexts with both cause and effect) was easier than V2 (i.e., item contexts with effect only). There is not much difference in preferences for the pair in D3 among ESLs, ELLs, and non-ELLs, between males and females, or among high/medium/low performers.

During the cognitive interview, most student responses (29 out of 36) selected the same option for the pair. If students chose the correct option for V1, they tended to choose the same correct option for V2. Among 21 who made the same responses on the both versions, 17 were correct on both versions, 6 were only correct on one version, and 10 were wrong on both versions. Later, I tracked those 36 students' responses on test items; 21 had the same options for the pair, and the other 15 did not have the same options for the pair.

The landscape of student interviews for D3 is provided in Table 24.

Table 24

Landscape of Student Interviews for D3: Sequence of Cause and Effect

Item	Item Difficulty Preference				
	V1 < V2 (n=23)		V1 = V2 (n=2)	V1 > V2 (n=7)	Mixed (n=4)
	Reason 1: V1 primed/provided students to consider the key information.	NA [a]	Reason 2: V1 and V2 were similar.	Reason 3: V2 was shorter and easier to understand.	Reason 4: Students did not understand the manipulation sentence.
LS_D3_L	1 ELL 1 non-ELL 2 ESLs	1 ELL 1 non-ELL	1 non-ELL	1 ELL 1 ESL	
LS_D3_H	3 ELLs 3 non-ELLs			2 ESLs	1 ESL
PS_D3_L	1 ELL 2 non-ELLs 2 ESLs		1 ESL	1 ELL	1 ELL 1 non-ELL
PS_D3_H	1 ELL 2 non-ELLs 2 ESLs	1 ELL		1 non-ELL 1 ESL	1 ELL
All items by ELLs	6	2	0	2	2
All items by non-ELLs	8	1	1	1	1
All items by ESLs	6	0	1	4	1

Note. [a] NA indicates the cases when the comments given were not related to the manipulation of the sequence dimensions.

I observed that students who preferred V1 (i.e., included both cause and effect) over V2 said that additional sentences explaining the cause of an effect primed or provided students clues for considering the key information. For example, one student explicitly stated that "If someone doesn't know what biodiversity means, the manipulation parts give them the background knowledge and help them to solve the problem." Another student explained why s/he preferred V1: "V1 tells how heat energy moves, and with that sentence I get a better understanding of the question. The manipulation helped [me] correctly answer the question."

Reasons that interviewed students reported for not preferring V1 (i.e., item contexts with both intention and action) are (1) students detected no obvious difference between V1 and V2; (2) students found V2 shorter and easier to understand; (3) students reported not understanding the manipulated sentence.

Many ELLs, non-ELLs and ESLs changed their item difficulty preferences to V1 (i.e., V1 is easier than V2) when the linguistic demands in item contexts increased for both life science and physical science items. Therefore, with the increase of linguistic demands in item contexts, adding a sentence to explain the cause of an effect provided some positive influence on students' perceptions of the item difficulty preference for V1.

4.4.4 Summary on Interviews

Overall, student cognitive interviews helped me better understand how students perceived sequence information presented in item contexts, thought on item contexts, and became clarified/confused by item contexts. Student cognitive interviews offered an opportunity to explore the results in depth beyond using psychometrics. First, it is very interesting to learn that student perceptions of the manipulation of context sequence differed from their test scores.

Although test scores indicate there was no significant difference between the manipulated version (V1) and the non-manipulated version (V2), more than half of interviewed students (i.e., 67% in D1, 56% in D2, and 64% in D3) commented that they thought V1 was easier than V2 for various reasons.

Second, when students did not like the manipulation of context sequence, they tended to give similar reasons across all three dimensions: (1) there was not much difference between two versions, so it is redundant to include the manipulation, and (2) the non-manipulated version (V2) with fewer words was easier, as it went straight to the point.

Third, the percentage of students being consistent with their responses for the pair in interviews differs from the percentage in two field tests. About 44%-58% of interviewed students were consistent with their responses on the pairs during two field tests, which was much lower than the percentage for the interview.

The results provide valuable insights into the effect that context sequence has influence on student performance, with implications for teaching, test writing, and testing research. In Chapter V, I will discuss these implications and make recommendations for future research.

Chapter V: Discussions and Conclusions

In this chapter, I highlight the major findings from the results section, and discuss the limitations that resulted from the research methods, and offer suggestions for further research. Last, I explain the implications of the major findings to testing research and educational practices.

5.1 Summary and Discussions

This study aimed to investigate how the characteristics of the sequence of contextual information presented in science items relates to student performance. Specially, I examined three dimensions of context sequence: *sequence of events* (D1), *sequence of intention and action* (D2), and *sequence of cause and effect* (D3). In the following section, I discuss every research question in turn and explain in plain terms what the statistical results mean.

5.1.1 Three Dimensions of the Context Sequence

The results from analyzing the standardized residual correlations suggested that *local independence* was achieved. *Local independence* occurs when correctly answering one version in the pair of items is conditionally independent of correctly answering the other version in the pair. One explanation for the finding of local independence in items in the present study could be that most of the students were medium/low performers with only partial understanding of science principles. Since each pair was assigned separately in two tests, through the benefits of the cross-over design of field tests used in this study, students did not remember the previous items and took the items in the second field test as new ones. Thus, one explanation is that student responses on items were inconsistent because of misconceptions about scientific

principles held by the majority of students, who were medium/low performers. Because of such inconsistent performance, *local independence* was achieved.

For D1: *sequence of events*, in opposition to my hypothesis, linearly ordered items without key words were significantly easier for ESLs than were those with key words for one pair of physical science items examining the relation between the distance from the light and the size of the shadow. After adding the key words, about 17% (i.e., 44 out of 263) more ESLs incorrectly answered the manipulated version (V1) compared to V2. In contrast to the other three pairs of items in D1, key words in this pair were added in two different paragraphs: "first" was added in the first paragraph before the diagram, and "second" was added in the second paragraph after the diagram. I speculate that the layout of manipulations jeopardized the connective textual relationships (Coirier, Favart, & Chanquoy, 2002), thus creating some mental barriers for the cognitive processing.

Comparing the pattern of two pairs of physical science items with the pattern of two pairs of life science items, the manipulated version (V1) in which sequential key words were added to item contexts in life science items showed promising results. This suggests that adding key words to item presentation facilitates to some extent student ability to comprehend the question being asked and thereby improves performance. One of the topics the two pairs of life science items used to test the *sequence of events* dimension examined the growth of an organism, and the other examined the matter and energy in organisms and ecosystems. An organism's genes interact with various environments throughout its lifespan, and according to previous research, when constructing an item to test this concept, it is helpful to use key words to identify the historical sequence of environments to which the developing organism is exposed (Griffiths,

2000). The results of the present study support these conclusions. In student cognitive interviews, many students suggested that adding sequential key words in item contexts would offer clues as to what happened next, and such additions helped them follow the steps of the investigation to answer the question. Considering different levels of linguistic demands in the two pairs of life science items, all interviewed ELL and non-ELL students preferred the manipulated version (V1), in which key words were added, over V2.

For D2: *sequence of intention and action*, it was hypothesized that if an intention is provided prior to action, items will be easier for students than when action is described without an intention being provided. Results showed that ESLs performed significantly better on the manipulated version (V1), in which an intention was added in item contexts in one pair of life science items examining structure, function, and information processing, as predicted. However, U.S. students (i.e., ELLs and non-ELLs) reported finding the added intention confusing. Their test results corroborate this, as it was found that U.S. students performed relatively less well on all four pairs of items to which intention was added. These items include two pairs of life science items examining structure, function, and information processing, and CO_2 concentration and photosynthesis, as well as two pairs of physical science items that examined energy transformations, heat, and temperature, and the boiling point of water at high altitudes.

Two explanations may help elucidate the statistical results showing that adding intention in item contexts significantly facilitate ESLs to achieve better scores on one particular life science item examining structure, function, and information processing. First, when adding the sentence "Corey wanted to know how adding a certain chemical to water would affect the temperature of water" to indicate the intention of this science investigation on the manipulated version (V1), this

sentence itself implied the meaning that the temperature of water would change. Based on this level of an implied meaning, ESLs could eliminate distractors A (i.e., The chemical always heats water to the same temperature.) and B (i.e., The temperature of water is not affected by the amount of the chemical.). After replacing this sentence with a descriptive sentence in V2, "Corey performed an experiment about the water temperature," about 15% of ESLs were distracted by other options. Yet this situation did not apply to ELLs and non-ELLs for this pair of life science items.

Second, it is possible that ESLs were less familiar than US students with the topics in the items because items were selected and developed based on the U.S. science curricula. And providing an explanation of the intention of the investigation may have helped ESLs narrow down the search database in their mind and quickly activate related scientific principles to answer the question (Bassok & Novick, 2012). U.S. students complained in interviews that they did not understand the manipulated sentences, which confused them because the students had a difficult time making connections between the ideas. This suggests that adding key words for students who already have a grasp of the material may merely complicate the question.

For D3: *sequence of cause and effect*, in opposition to my hypothesis, item contexts with only an effect presented were significantly easier for ESLs than those that contain both a cause and an effect for one pair of physical science items. This pair of physical science items examined the conservation of energy and energy transfer. In the manipulated version (V1), a sentence indicating the theory of heat transfer, "Heat energy moves from warmer objects to colder ones," was added in to facilitate problem solving. In contrast to expectations, after adding in this sentence (cause), the number of students who chose distractor C (i.e., 70 ºC and 25 ºC)

doubled. Student cognitive interviews clearly revealed the reasons why distractor C became more appealing to ESLs. One interviewed ESL said, "Because this item (PS_D3_V1_L2) says the heat energy moves from warmer objects to colder objects, 90 ºC's coffee moves some heat to the room. So the temperature of coffee becomes lower than 90 ºC but higher than the room temperature 20 ºC. The room temperature should go up to more than the current 20 ºC. The 5 ºC's water should become warmer than the 20 ºC. So I choose option C: the coffee becomes 70 ºC and the water becomes 25 ºC." Student cognitive interviews provided evidence that those students who did not fully understand the theory of heat transfer could misinterpret the manipulated sentence differently from its intended meaning. In this case, causal mechanisms that explain how mechanisms work (Graesser et al., 2003) did not facilitate problem solving well and instead jeopardized the reasoning process, thus leading to poorer student scores.

In all four pairs of science items from the perspectives of descriptive comparisons, the manipulated version (V1), in which a cause was added in item contexts to indicate the theory of a scientific phenomenon, did not help non-ELLs at all, and barely helped ESLs and ELLs. Two pairs of life science items examined natural selection and adaptations; and the food web. Two pairs of physical science items examined temperature changes the mass of an object; and the energy transfer. In student cognitive interviews, some non-ELLs mentioned that manipulation was not needed, or they preferred a shorter version of item contexts. Specifically, non-ELLs, especially high performers with solid science knowledge, reported that they considered the manipulated sentence as redundant information. This reflects results found in D2, suggesting that for those students with a solid grasp of the material, any additional information – even that meant to give context – may interfere with, rather than aid, comprehension.

In sum, three dimensions of context sequence have different impacts on different demographic student groups. Although literature from cognitive psychology suggests that adding key words, providing an intention, and stating a cause could facilitate the problem solving process and reduce the cognitive load (Bassok & Novick, 2012; Graesser et al., 2003; Mayer &Moreno, 2003), it does not necessarily lead to better student performance in test scores.

5.1.2 Discussions on DIF Sources

Differential item functioning (DIF) occurs when people from different groups with the same underlying true ability have a different probability of giving a certain response on a test item (Embretson & Reise, 2013). DIF analysis provides an indication of unexpected behavior of items on a test. In the present study, I performed two DIF runs (i.e., one DIF run for ELLs and non-ELLs, another DIF run for ESLs and non-ELLs). One pair of life science items was flagged as DIF items from D1: *sequence of events*, which tested students on how to improve an experiment so that the conclusion (i.e., that certain seeds require sunlight to germinate) was strengthened. Two pairs of physical science items were flagged as DIF items from D2: *sequence of intention and action*. One pair examined the relation of boiling temperature to air pressure or altitude, and the other pair examined student understanding of controlled and uncontrolled variables in an experiment aimed at testing which of two sources heats water faster. This suggests that either adding the sequential key words or adding an intention in item contexts tends not be the source of DIF, because both versions of items from D1 and D2 were flagged as DIF items.

Interestingly, in D3: *sequence of cause and effect*, only the manipulated version (V1, item contexts with both cause and effect) was identified as a DIF item in two DIF runs. This may

indicate that adding a cause in item contexts can be a source of DIF. Different sub-groups of students may interpret the manipulated sentence in different ways; thus, the manipulated sentence may help or hinder different groups of students' answering process. For instance, the flagged physical science item (Temperature changes the mass of an object) in the manipulated version (V1) favored the reference group (i.e., non-ELLs) in two DIF analyses. A comparison of student cognitive interviews with the DIF results suggests that if students could understand the underlying scientific information embedded in the manipulated sentence, the manipulated sentence would become an important clue to activate student answering process (Ahmed & Pollitt, 2007; Bassok & Novick, 2012). This finding confirmed Ahmed and Pollitt's study, which found that if the manipulation activates critical underlying scientific concepts tapped in the question, item contexts are focused to facilitate the problem solving process. Otherwise, the manipulation could add in unfocused information in item contexts, thus interfering with the comprehension.

Four science items (i.e., 3 life science and 1 physical science) were twice flagged as DIF items in two DIF runs, three of which favored non-ELLs and ESLs (i.e., 3 life science items). One of the life science items focused on structure, function, and information processing. One pair of the life science items focused on the growth of organisms.

There was one physical science item (PS_D3_V1_L) twice flagged as a DIF item favoring the reference group (i.e., non-ELLs) in two DIF runs. This physical science item examined student understanding of how temperature changes the state of water. I speculate that the manipulated sentence, "Temperature changes the state of water," helped non-ELLs to activate the related science principles to decode the item. Thus, non-ELLs could perform better on this

item than on the version without manipulation. Specifically, the underlying concept, that different states of water have different shapes and volumes, was tested in this item. Frozen water has a definite shape and volume. The particles in a solid are very close to one another and are in fixed positions, making solids relatively incompressible. The particles in a liquid are not in fixed positions and have a greater freedom to move than in a solid. Interestingly, this physical item was the easiest item for non-ELLs. Also, I found this item was one of the easiest items for non-ELLs in the test.

5.1.3 Student Cognitive Interviews

Student cognitive interviews offered an approach to assist researchers in understanding how students process information in item contexts. First, it is very interesting to learn that the information students provided during the interviews on the manipulation of context sequence differed from test scores. Although many interviewed students indicated that they preferred (i.e., found easier to understand) the manipulated version, student test scores indicated that students actually did slightly better on the version that was not manipulated. There are three factors as plausible explanations: (1) under the intense conditions of field tests in classrooms, students may not notice minor changes to items, such as adding a few words or a sentence. When we presented the pairs one by one (see Appendix C) in the less stressful student interviews, many students could notice the minor changes. (2) Student interviews provided snapshots of students' perceptions on test items. Some students were not familiar with the think-aloud strategy, so interviewers had to frequently remind students to verbalize what they were thinking. Thus, students' perceptions might not be fully captured in this study. (3) After canceling out

measurement errors from the test scores, student preferences on context sequence were not necessarily translated into test scores.

Second, students gave similar reasons to explain why they thought the manipulation of context sequence was easier for them. Students said the non-manipulated version (V2) was shorter and easier to understand. This results suggests the manipulated information in item contexts is irrelevant and unfocused to the problem solving (Ahmed & Pollitt, 2007; Bassok & Novick, 2012). I speculate that any redundant information added to surface features of the problems could potentially interfere with the student answering process.

Third, the percentage of students who were consistent with their responses for the pair in interviews was much higher than the percentage in two field tests. To be more specific, results indicated a 41% difference in D1, 39% difference in D2, and 22% difference in D3. Compared to low performers, high and medium performers tended to be more consistent with their responses in tests. This finding is consistent with Al-Hamly and Coombe's study (2005) that found high performers were less likely to change answers, compared with low performers. In the present study, many low performers did not perform consistently on the pairs between the first and second test, likely because they had only partial understanding of the concepts being tested, and their misconceptions were easily captured through changing responses on test items.

5.2 Limitations and Recommendations

This study has several key limitations that preclude overgeneralization and warrant further research. First, a major limitation was the small sample size of items that were tested in this study. Constrained by the resources needed for the item development and also a reasonable testing length for the pencil-paper test, this study included only 12 pairs of multiple-choice items,

resulting in relatively low power for data analyses. The number of items per topic needs to be expanded in future studies to increase the power of analysis. As the power increases, the chances of a Type II error (false negative) decreases (Ross, 2009).

Second, minor changes in item contexts may not have a significant impact on student performance. From the student cognitive interview, I learned that some students required prompting from the interviewer to notice the minor difference between the pair, and some students could not easily distinguish the differences between the two questions in the pair. This suggests that there is a need for considering at least several manipulations in item contexts, rather than only one, as in the present study. In a previous study by Abedi and Lord (2001), researchers made several modifications to reduce linguistic demands in math items, and found that the linguistically modified version had a statistically significant impact on student performance.

A third limitation is the use of the 1PL model in the study. In the present study, this model was chosen because most original released item sources (such as NAEP, TIMSS, PISA) used a 1PL model and/or CTT. In a future study with a larger sample size of items, it would be possible to integrate the explanatory variables into the item response modeling: explanatory item response models (EIRM; De Boeck & Wilson, 2004; Briggs, 2008). This approach would also provide a more powerful way to investigate differential item functioning between the reference group and the focal group. EIRM is a multilevel model, having both a within-person level and a between-person level, with the multidimensional Rasch Model characterizing the former, and a population model (i.e., latent regression) characterizing the latter. Such EIRM-based analysis would likely be more powerful than the two-step approach (i.e., basic IRT modeling followed by analysis of estimated parameters and fit statistics) used in the present study.

The last limitation has to do with the fact that the ELL students were considered as a homogeneous group due to the small sample size. In this study, ELLs were a mixture of White, Asian, Latino, African American, and Multiracial students. Less than 10 ELLs fell into the same ethnic group due to the small ELL sample size in the present study. It was impossible to divide those 103 ELLs into several different ELL groups, and still conduct statistical analyses with enough power, and generate research findings. Oliveri, Ercikan, and Zumbo (2014) demonstrated that ethnic group is an important variable to consider when studying ELLs, as it relates to factors such as English proficiency, which can affect student comprehension and test-taking abilities. Thus, future research needs to take into account the fact that ELLs' English proficiency levels can range widely. Heterogeneity in ELL English proficiency may be tied to various factors. As Oliveri, Ercikan, and Zumbo (2014) documented, linguistic heterogeneity within ELLs can occur for a number of reasons, including different lengths of time residing in English-speaking countries, degrees of exposure to English-speaking environments, and amounts of English instruction. Moreover, Solano-Flores (2011) argued that an item with certain linguistic features (e.g., notation) may be effective in minimizing measurement errors due to language proficiency for some, but not all ELL students. The language background such as cultural anthropology and language development might play an important role in how students make sense of an item and construct knowledge (e.g., Bialystok, 2002; Vygotsky, 1978; Wertsch, 1985). Thus, to further the results of the present study, future work must take into consideration the heterogeneity of ELLs and explore how and why sequence dimension may cause potential bias for some but not other ELL sub-groups.

5.3 Implications

This exploratory study has generated a number of implications that may be of interest to practitioners, item writers, educators, and researchers. It should be stressed that the implications presented are by no means exhaustive. They are, however, intended to stimulate thinking on how the insights from this study might impact education in a very broad way.

In this dissertation, I used a systematic approach to unfold and characterize the sequence of contextual information used in science item contexts. Sequence of contextual information refers to the order of descriptions involved in the item contexts. Three dimensions were examined in this study: sequence of events (D1), sequence of intention and action (D2), and sequence of cause and effect (D3). A summary of the study results offered greater clarity in regard to the impacts of sequence of contextual information on performance in different student demographics.

Practitioners could utilize those student preferences (that correlate with test scores) indicated in the cognitive interviews to incorporate context sequence in their classroom teaching. For instance, when introducing an investigation to students, teachers could introduce the steps by specifying the order, stating the intention of the investigation, and re-emphasizing the theory involved in the investigation.

Item writers might find benefit in considering the characteristics of item contexts while developing contextualized items. For instance, when writing assessment items for Chinese students who have limited English proficiency and also have cultural backgrounds different from that of U.S. students, item writers could consider adding in an intention to help those students to present the purposes of an investigation, which may help student performance, as suggested by the results of the present study.

Recommendations for future study include several points. First, it is recommended that, as in the present study, future research follow systematic guidelines to examine characteristics of item contexts. Second, to improve results, it is recommended that researchers use several manipulations on item contexts instead of only one. Last, student interviews such as the ones performed in this study could help researchers to deepen their understanding of how variations in item contextualization impact student performance.

References

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

Abedi, J. (2006). Language issues in item-development. In S. M. Downing, & T. M. Haladyna. (Eds.). *Handbook of Test Development* (pp. 377-298). New Jersey: Lawrence Erlbaum Associates, Publishers.

Aguirre-Muñoz, Z., & Baker, E. L. (1998). *Improving the equity and validity of assessment-based information systems* (CSE Tech. Rep. No. 462). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Ahmed, A., & Pollitt, A. (1999). *Curriculum demands and question difficulty*. Paper presented at the International Association of Educational Assessment Conference (IAEA), Slovenia.

Ahmed, A., & Pollitt, A. (2001). *Improving the validity of contextualized questions.* Paper presented at the British Educational Research Association Annual Conference, Leeds.

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.

Ahmed, A., & Pollitt, A. (2010). The support model for interactive assessment. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 133-167.

Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing, 22*(4), 509-531.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). *White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS).* Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1999). *Applications and misapplications of cognitive psychology to mathematics education*. Retrieved from http://act-r.psy.cmu.edu/papers/misapplied.html

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Barber, M. (2000). A comparison of NEAB and Salters A-level chemistry: Student views and achievements. *Unpublished MA thesis, University of York, York.*

Bassok, M. & Novick, L. R (2012). Problem Solving. In K. J. Jolyoak & R.G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning* (Chapter 21, 413-432). New York, NY: Oxford University Press.

Beasley, W., & Butler, J. (2002). *Implementation of context-based science within the freedoms offered by Queensland schooling*. Paper presented at the annual meeting of Australian Science and Education Research Association Conference, Townsville, Queensland, July.

Beitzel, B. D., Staley, R. K., & DuBois, N. F. (2011). The (in) effectiveness of visual representations as an aid to solving probability word problems. *Effective Education*, *3*(1), 11-22.

Bellocchi, A., King, D. T., & Ritchie, S. M. (2011). Assessing students in senior science: an analysis of questions in contextualised chemistry exams. In *Proceedings of the 1st International Conference of STEM in Education 2010*. Science, Technology, Engineering and Mathematics in Education Conference.

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

Bialystok, E. (2002). Cognitive processes of L2 users. In V. J. Cook (Ed.), *Portraits of the L2 user* (pp. 145–165). Clevedon, UK: Multilingual Matters.

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.

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.

Bransford, J., Brown, A. L., Cocking, R. R., & National Research Council (U.S.). (2000). *How people learn: Brain, mind, experience, and school*. Washington, D.C: National Academy Press.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89-118.

Brown, F. G. (1970). *Principles of educational and psychological testing*. Hinsdale, Ill: Dryden 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.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293-332.

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *The British Journal of Educational Psychology, 62*(2), 233-246.

Chipman, S. F., Marshall, S. P., & Scott, P. A. (1991). Content effects on word problem performance: A possible source of test bias?. *American Educational Research Journal*, *28*(4), 897-915.

Clausen-May, T. (2005). *Teaching maths to pupils with different learning styles*. London: Paul Chapman Publishing.

Clausen-May, T. (2006). Reality and context in maths test questions. *Mathematics in School*, *35*(5), 9-11.

Clement, J. (1982). Algebra word problem solutions: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education*, *13*(1), 16-30.

Coffey, J., & Alberts, B. (2013). Improving education standards. *Science*, *339*(6119), 489-489.

Coirier, P., Favart, M., & Chanquoy, L. (2002). Ordering and structuring ideas in text: From conceptual organization to linguistic formulation. *European Journal of Psychology of Education*, *17*(2), 157–175.

Coştu, B. (2007). Comparison of students' performance on algorithmic, conceptual and graphical chemistry gas problems. *Journal of Science Education and Technology*, *16*(5), 379-386.

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.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, *42*(3), 475-493.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test*. Research Report 11–26. Princeton, NJ: Educational Testing Service.

De Boeck, P. and Wilson, M., eds. (2004). *Explanatory item response models: a generalized linear and nonlinear approach.* New York: Springer.

De Bock, D., Verschaffel, L., Janssens, D., Van Dooren, W., & Claes, K. (2003). Do realistic contexts and graphical representations always have a beneficial impact on students' performance? Negative evidence from a study on modelling non-linear geometry problems. *Learning and Instruction*, *13*(4), 441-463.

Dennis, I., Handley, S., Bradon, P., Evans, J., & Nestead, S. (2002). Approaches to modeling item-generative tests. In S. H. Irvine, & P. C. Kyllonen (Eds.). *Item generation for test development*, (pp. 53 –72). Mahwah, NJ: Erlbaum.

Duncker, K. (1945). On problem-solving (L. S. Lees, Trans.). *Psychological Monographs*, *58* (Whole No. 270). (Original work published 1935).

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education*, *15*(1), 49-74.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (revised ed.). Cambridge MA: MIT Press.

Evans, J. S. B. T., Newstead, S. E., Byrne, R. M. J., & Foos, P. W. (1996). Human Reasoning: The Psychology of Deduction. *Contemporary Psychology, 41*(9), 917.

Ferrara, S., Duncan, T. G., Freed, R., Velez-Paschke, A., McGivern, J., Mushlin, S., Mattessich , A., Rogers, A., & Westphalen, K. (2004). *Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment*. In Annual Meeting of the American Educational Research Association, San Diego, *CA*.

Finkenbinder, E. O. (1913). The curve of forgetting. *The American Journal of Psychology*, 8-32.

Gabora, L., & Kitto, K. (2013). Concept Combination and the Origins of Complex Cognition. In *Origins of Mind* (pp. 361-381). Springer Netherlands.

Garvin, P. L. (1973). Linguistics as a Resource in Language Planning.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking reading comprehension*, 82-98.

Grawe, N. D. (2011). Beyond math skills: Measuring quantitative reasoning in context. *New Directions for Institutional Research*, *2011*(149), 41-52.

Griffiths AJF, Miller JH, Suzuki DT, et al. (2000). An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman. Genes, the environment, and the organism. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21842/

Gorin, J. S., & Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394-411.

Gutierrez, J. R. M., & Ikeda, H. (2009). Response pattern analysis on the burning candle experiment: TIMSS-based study.

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

Hambleton, R., & Russell, J. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, J. (1991). Fundamentals of Item Response Theory. London: Sage Publications.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.

Heredia, S., Furtak, E. M., & Morrison, D. (2012). *Item context: how organisms used to frame natural selection items influence student response choices*. In Proceedings of the National Association for Research in Science Teaching (NARST) annual conference. Indianapolis, IN.

Hickendorff, M. (2013). The Effects of Presenting Multidigit Mathematics Problems in a Realistic Context on Sixth Graders' Problem Solving. *Cognition and Instruction*, *31*(3), 314-344.

Hoey, Michael. 2005. *Lexical Priming. A new theory of words and language.* London: Routledge.

Hofstein, A., Kesner, M., & Ben-Zvi, R. (1999). Student perceptions of industrial chemistry classroom learning environments. *Learning Environments Research*, *2*(3), 291-306.

Hogg, M. A., & Cooper, J. (2003). *The Sage handbook of social psychology*. Sage.

Holland, P. W., & Thayer, D. T. (1985). An alternate definition of the ETS delta scale of item difficulty. *ETS Research Report Series*, *1985*(2), 1-10.

Holland, P. W., & Thayer, D. T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer, and H. I. Brown (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ilich, M. O. (2013). *Differential Item Functioning (DIF) among Spanish-Speaking English Language Learners (ELLs) in State Science Tests*. University of Washington at Seattle, Educational Psychology Department, PhD Thesis.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2006). *Do children need concrete instantiations to learn an abstract concept*. In Proceedings of the 28th annual conference of the Cognitive Science Society (pp. 411-416).

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). Learning theory: The advantage of abstract examples in learning math. *Science*, *320*(5875), 454-455.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331-358.

Karabenick, S. A. (2006). *Cognitive validity and the construct validation of motivation-related assessments: What, why, and how*. In Annual Meeting of the American Educational Research Association, San Francisco, CA.

Kastberg, S. E., D'Ambrosio, B., McDermott, G., & Saada, N. (2005). Context matters in assessing students' mathematical power. *For the Learning of Mathematics*, *25*(2), 10-15.

Kelly, V. L. (2007). *Alternative assessment strategies within a context-based science teaching and learning approach in secondary schools in Swaziland*. Doctoral dissertation, Faculty of Education, University of the Western Cape.

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.

Li, M., Ruiz-Primo, M.A., Wills, K., & Giamellaro, M. (2012). *Instructionally sensitive assessments across three science units.* In American Educational Research Association Annual Meeting, Vancouver, B.C.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*,*22*(3), 249-264.

Linacre, J. M. (2009). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of Applied Measurement*, *10*(2), 2-13.

Linacre, J. M., & Wright, B. D. (2000). Winsteps. *Computer software. Chicago: MESA Press*.

Lehman, S., & Schraw, G. (2002). Effects of coherence and relevance on shallow and deep text processing. *Journal of Educational Psychology*, *94*(4), 738-750.

Leighton, J.P. & Gokiert, R.J. (2005). *The Cognitive Effects of Test Item Features: Informing Item Generation by Identifying Construct Irrelevant Variance*. In Annual Meeting of the National Council on Measurement in Education (NCME), Montreal, Quebec, Canada.

Little, C., & Jones, K. (2010). The effect of using real world contexts in post-16 mathematics questions. *British Society for Research into Learning Mathematics (BSRLM)*: 137-144.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, *78*(2), 333-368.

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. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43-52.

Messick, S. (1993). Validity. In R. L. Linn (Ed), *Educational measurement* (2nd ed., pp. 13-104). Phoenix: American Council on Education and Oryx Press.

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.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Moreno, R., Ozogul, G., & Reisslein, M. (2011). Teaching with concrete and abstract visual representations: Effects on students' problem solving, problem representations, and learning perceptions. *Journal of educational psychology*, *103*(1), 32-47.

Newstead, S., Bradon, P., Handley, S., Evans, J., & Dennis, I. (2002). Using the Psychology of Reasoning to Predict the Difficulty of Analytical Reasoning Problems. In S. H. Irvine, & P. C. Kyllonen (Eds.). *Item generation for test development*, (pp. 35 –52). Mahwah, NJ: Erlbaum.

NGSS. (2013). "*Next Generation Science Standards*: For States, By States," online at <http://bit.ly/Z2V8YS>.

Nickson, M., S. Green. (1996). *A study of the effects of context in the assessment of the mathematical learning of 10/11 year olds*. British Educational Research Association Annual Conference.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). “I never thought of it as freezing”: How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778-803.

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286-300.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question?. *Behavior Research Methods*, *40*(4), 1001-1015.

Perin, D. (July 01, 2011). Facilitating student learning through contextualization: A review of evidence. *Community College Review, 39*(3), 268-295.

Piantadosi, S. (2005). Crossover designs. *Clinical trials: a methodologic perspective, Second Edition,* 515-527.

Pollitt, A., & Ahmed, A. (1999). *A new model of the question answering process*. Paper presented at the International Association of Educational Assessment Conference (IAEA) conference, Slovenia.

Pollitt, A., & Ahmed, A. (2000). *Comprehension failures in educational assessment*. In European Conference on Educational Research, Edinburgh.

Pollitt, A., Marriott, C., & Ahmed, A. (2000). *Language, contextual and cultural constraints on examination performance*. In International Association for Educational Assessment Conference.

QSR International. (2008). NVivo qualitative data analysis software (Version 8).

Reisslein, M., Moreno, R., Ozogul, G. (2010). Pre-college electrical engineering instruction: The impact of abstract vs. contextualized representation and practice of learning. *Journal of Engineering Education*, *99*(3), 225–235.

Rivet, A. E., & Krajcik, J. S. (2008). Contextualizing instruction: Leveraging students' prior knowledge and experiences to foster understanding of middle school science. *Journal of Research in Science Teaching*, *45*(1), 79-100.

Roth, W. M., & Hwang, S. (2006). On the relation of abstract and concrete in scientists' graph interpretations: A case study. *The Journal of Mathematical Behavior*, *25*(4), 318-333.

Ruiz-Primo, M. A. & Li, M. (2012). The impact of item context on students' performance: the case of the 2006 and 2009 PISA science. Manuscript submitted to *Teachers College Record*.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships Among Item Cognitive Complexity, Contextual Demands, and Item Difficulty: Implications for Achievement-Level Descriptors. *Educational Assessment*, *18*(2), 99-121.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Seiler, S. (2012). *The impact of item characteristics on contextualized personality assessment*. Doctoral dissertation. University of Illinois at Urbana-Champaign, Psychology Department, PhD Thesis.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126.

Siegel, M.A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching*, *44*(6), 864–881.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In *Generating Items for Cognitive Tests: Theory and Practice., Nov, 1998, Educational Testing Service, Princeton, NJ, US; This chapter was presented at the aforementioned conference.*. Lawrence Erlbaum Associates Publishers.

Smith, F. (2004). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Routledge, Psychology Press.

Solano-Flores, G. (2011). Language issues in mathematics and the assessment of English language learners. In K. Tellez, J. Moschkovich, & M. Civil (Eds.), *Latinos/as and mathematics education: Research on Learning and Teaching in Classrooms and Communities,* (pp. 283-314). Charlotte, NC: Information Age Publishing.

Solano-Flores, G., Barnett-Clarke, C., & Kachchaf, R. R. (2013). Semiotic Structure and Meaning Making: The Performance of English Language Learners on Mathematics Tests. *Educational Assessment*, *18*(3), 147-161.

Solano-Flores, G., & Wang, C. (2011). *Conceptual framework for analyzing and designing illustrations in science assessment: Development and use in the testing of linguistically and culturally diverse populations*. In the Annual Conference of the National Council on Measurement in Education, New Orleans, LA.

Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Taber, K. S. (2003). Examining structure and context-questioning the nature and purpose of summative assessment.

Taylor, C. S., & Nolen, S. B. (1996). A contextualized approach to teaching teachers about classroom-based assessment. *Educational Psychologist*, *31*(1), 77-88.

Turmo, A., & Elstad, E. (2009). What factors make science test items especially difficult for students from minority groups?. *Nordic Studies in Science Education*, *5*(2), 158-170.

Van Den Heuvel-Panhuizen, M. (2005). The role of contexts in assessment problems in mathematics. *For the Learning of Mathematics*, *25*(2), 2-23.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wang, T., Li, M., Thummaphan, P., & Ruiz-Primo, M. A. (2013). *The link between sequence of item context and students' performance in science assessment.* Paper presented at the Annual Meeting of National Association for Research in Science Teaching (NARST), Puerto Rico.

Wang, T., Liaw, Yuan-Ling., Li, M., & Taylor, C. (2014). *Identifying Science Item Context Characteristics for English Language Learners (ELLs) and non-ELLs by Differential Item Functioning (DIF).* Paper presented at the Annual Meeting of American Educational Research Association (AERA), Philadelphia, PA.

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.

Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds). *Handbook of test development* (pp.303-327). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind.* Cambridge, MA: Harvard University Press.

Wiliam, D. (1997). Relevance as MacGuffin in mathematics education. *British Educational Research Association*. York.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*(3-4),139-159.

Wolff, J. S. (1996). A study of the effect of context and test method in evaluating safety symbols.

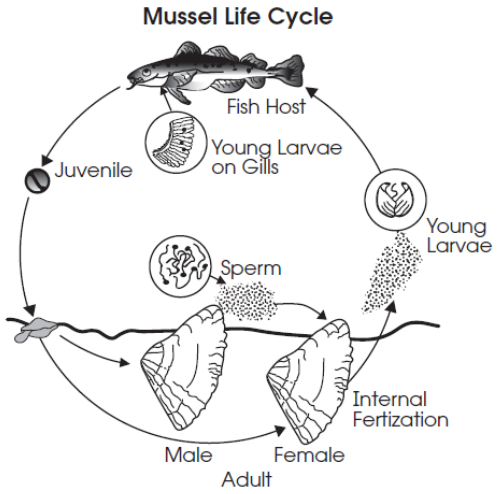
Wright, B. D., & Stone, M. H. (1979). *Best Test Design. Rasch Measurement.* Chicago, IL: MESA Press.

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. In P. W. Holland, and H. Wainer (Eds.), *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Appendix A: Original Items and Testing Items in V1 and V2

Original Released Item	Version 1	Version 2
LS_D1_L Kim wanted to determine if certain seeds require sunlight to germinate. She placed one seed in a moist paper towel in the sunlight and another seed in an equally moistened paper towel in a dark closet. The seed in the sunlight germinated but the one in the closet did not. Kim reported to the class that this type of seed needs sunlight in order to germinate. Given this information, which of the following would best describe an improvement in Kim's experiment that would strengthen her claim? A. Use many seeds to conduct the experiment. B. Start the samples on different days. C. Use different amounts of water. D. Place the seeds in new locations. Item source: Michigan YR2008 Gr8	LS_D1_V1_L1 Kim wanted to determine if certain seeds require sunlight to germinate. First, she placed one seed in a moist paper towel in the sunlight. Then, she placed another seed in an equally moistened paper towel in a dark closet. After a few days, the seed in the sunlight germinated but the one in the closet did not. Last, Kim reported to the class that this type of seed needs sunlight in order to germinate. Given this information, which of the following would best describe an improvement in Kim's experiment that would strengthen her claim? A. Use many seeds to conduct the experiment. B. Start the samples on different days. C. Use different amounts of water. D. Place the seeds in new locations.	LS_D1_V2_L1 Kim wanted to determine if certain seeds require sunlight to germinate. She placed one seed in a moist paper towel in the sunlight. She placed another seed in an equally moistened paper towel in a dark closet. The seed in the sunlight germinated, but the one in the closet did not. Kim reported to the class that this type of seed needs sunlight in order to germinate. Given this information, which of the following would best describe an improvement in Kim's experiment that would strengthen her claim? A. Use many seeds to conduct the experiment. B. Start the samples on different days. C. Use different amounts of water. D. Place the seeds in new locations.
LS_D1_H Male mussels release sperm into the water. Female mussels take the sperm into their gill chambers where fertilization occurs. Young mussel larvae are released into the water where they float freely until	LS_D1_V1_L2 First, male mussels release sperm into the water. Then female mussels take the sperm into their gill chambers where fertilization occurs. Later, young mussel larvae are released into the water where	LS_D1_V2_L2 Male mussels release sperm into the water. Female mussels take the sperm into their gill chambers where fertilization occurs. Young mussel larvae are released into the water where they float freely until

they attach to the gill of a host fish. After a few weeks, they reach the juvenile stage and drop off. After the juvenile drops off the fish gill, it burrows into the river bed and begins the life cycle all over again.



Source: Department of Fisheries and Wildlife, Virginia Tech

The parasitic behavior of the larvae benefits the mussel in two ways. One benefit is that the fish provides nutrition for the larvae when they are attached to its gill.

What is the second way this behavior enhances the survival of the mussel species?

they float freely until they attach to the gill of a host fish. After a few weeks, young mussel larvae reach the juvenile stage and drop off.

The parasitic behavior of the larvae benefits the mussel in two ways. One benefit is that the fish provides nutrition for the larvae when they are attached to its gill.

What is the second way this behavior enhances the survival of the mussel species?

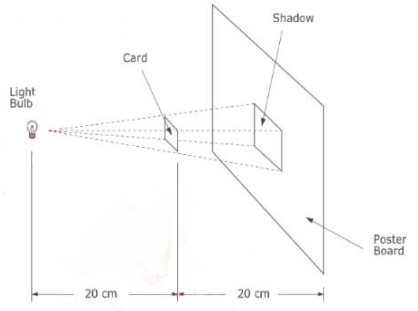
A. The large size of the fish provides the mussel larvae with plenty of room to grow.
B. The mobility of the fish spreads the mussels to areas they would otherwise be unable to reach.
C. The parasitism increases the opportunity for the mussels to mate with other mussel species.
D. The location of the larvae on the gills of fish reduces the exposure of the larvae to oxygen-rich water.

they attach to the gill of a host fish. A few weeks, they reach the juvenile stage and drop off.

The parasitic behavior of the larvae benefits the mussel in two ways. One benefit is that the fish provides nutrition for the larvae when they are attached to its gill.

What is the second way this behavior enhances the survival of the mussel species?

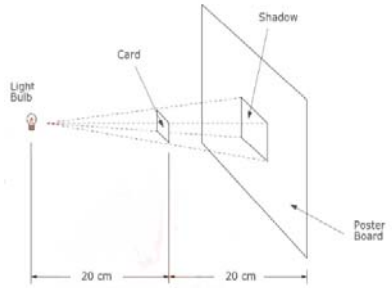
A. The large size of the fish provides the mussel larvae with plenty of room to grow.
B. The mobility of the fish spreads the mussels to areas they would otherwise be unable to reach.
C. The parasitism increases the opportunity for the mussels to mate with other mussel species.
D. The location of the larvae on the gills of fish reduces the exposure of the larvae to oxygen-rich water.

A. The large size of the fish provides the mussel larvae with plenty of room to grow. B. The parasitism increases the opportunity for the mussels to mate with other mussel species. C. The mobility of the fish spreads the mussels to areas they would otherwise be unable to reach. D. The location of the larvae on the gills of fish reduces the exposure of the larvae to oxygen-rich water. Item source: Ohio YR2007 Gr8		
PS_D1_L A tiny light bulb was held 20 centimeters to the left of a square card, which was in turn held 20 centimeters to the left of a poster board, as shown. The shadow of the card on the poster board had a side of 10 centimeters.  	PS_D1_V1_L1 Lily conducted an investigation on the light. First, a tiny light bulb was held 20 centimeters to the left of a square card, which was in turn held 20 centimeters to the left of a poster board, as shown. The shadow of the card on the poster board had a side of 10 centimeters.	PS_D1_V2_L1 Lily conducted an investigation on the light. A tiny light bulb was held 20 centimeters to the left of a square card, which was in turn held 20 centimeters to the left of a poster board, as shown. The shadow of the card on the poster board had a side of 10 centimeters.

If the poster board is moved 40cm further to the right so that it is 80cm from the light, what will be the new side of the card's shadow on the poster board?

A. 5cm
B. 10cm
C. 15cm
D. 20cm

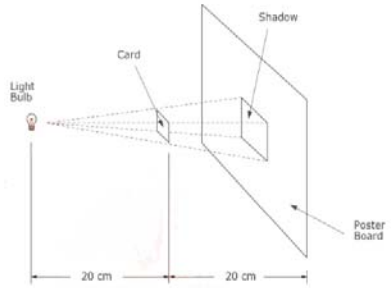
Item source: TIMSS YR2003 Gr8



Second, she moved the poster board 40cm further to the right. So the poster board was 80cm from the light and 60cm from the card.

What will be the new side of the card's shadow on the poster board?

A. 20cm
B. 15cm
C. 10cm
D. 5cm



She moved the poster board 40cm further to the right. So the poster board was 80cm from the light and 60cm from the card.

What will be the new side of the card's shadow on the poster board?

A. 20cm
B. 15cm
C. 10cm
D. 5cm

PS_D1_H

The table below contains two different recipes for cosmetics you can make yourself.

The lipstick is harder than the lip gloss, which is soft and creamy.

Lip gloss Ingredients:	**Lipstick Ingredients:**

PS_D1_V1_L2

The table below contains a recipe for lipstick you can make yourself. Oils and waxes are substances that will mix well together. Oils do not readily mixed with water, and waxes are not soluble in water.

Ingredients:
5 g castor oil

PS_D1_V2_L2

The table below contains a recipe for lipstick you can make yourself. Oils and waxes are substances that will mix well together. Oils do not readily mixed with water, and waxes are not soluble in water.

Ingredients:
5 g castor oil

5 g castor oil 0.2 g beeswax 0.2 g palm wax 1 teaspoon of colouring substance 1 drop of food flavouring **Instructions:** Heat the oil and the waxes in a container placed in hot water until you have an even mixture. Then add the colouring substance and the flavouring, and mix them in.	5 g castor oil 1 g beeswax 1 g palm wax 1 teaspoon of colouring substance 1 drop of food flavouring **Instructions:** Heat the oil and the waxes in a container placed in hot water until you have an even mixture. Then add the colouring substance and the flavouring, and mix them in.	1 g beeswax 1 g palm wax 1 teaspoon of coloring substance 1 drop of food flavoring **Instructions**: First, heat the oil and the waxes in a container placed in hot water until you have an even mixture. Then add the coloring substance and the flavoring, and mix them in.	1 g beeswax 1 g palm wax 1 teaspoon of coloring substance 1 drop of food flavoring **Instructions**: Heat the oil and the waxes in a container placed in hot water until you have an even mixture. Add the coloring substance and the flavoring, and mix them in.
Oils and waxes are substances that will mix well together. Oils cannot be mixed with water, and waxes are not soluble in water. Which one of the following is most likely to happen if a lot of water is splashed into the lipstick mixture while it is being heated?		Which one of the following is most likely to happen if a lot of water is splashed into the lipstick mixture while it is being heated? A. The mixture becomes firmer. B. The mixture is hardly changed at all. C. A creamier and softer mixture is produced. D. Fatty mixture floats on the water.	Which one of the following is most likely to happen if a lot of water is splashed into the lipstick mixture while it is being heated? A. The mixture becomes firmer. B. The mixture is hardly changed at all. C. A creamier and softer mixture is produced. D. Fatty mixture floats on the water.

A. A creamier and softer mixture is produced.
B. The mixture becomes firmer.
C. The mixture is hardly changed at all.
D. Fatty lumps of the mixture float on the water.

Item source: PISA YR2006 Gr8 S470Q02

LS_D2_L

Corey found that when he added a certain chemical to water, the water would heat up. He then performed an experiment in which he mixed different amounts of the chemical with water in a test tube and measured the temperature of the water.

The results of his experiment are shown in the table below.

Trail	Amount of Water	Amount of Chemical	Temperature of Water before Adding Chemical	Temperature of Water 2 Minutes after Adding Chemical
1	100 mL	0 grams	21 ºC	21 ºC
2	100 mL	5 grams	21 ºC	27 ºC
3	100 mL	10 grams	21 ºC	32 ºC

LS_D2_V1_L1

Corey wanted to know how adding a certain chemical to water would affect the temperature of water. He mixed different amounts of the chemical with water in a test tube and measured the temperature of the water. The results of his experiment are shown in the table below.

Trail	Amount of Water	Amount of Chemical	Temperature of Water before Adding Chemical	Temperature of Water 2 Minutes after Adding Chemical
1	100 mL	0 grams	21 ºC	21 ºC
2	100 mL	5 grams	21 ºC	27 ºC
3	100 mL	10 grams	21 ºC	32 ºC
4	100 mL	15 grams	21 ºC	35 ºC

LS_D2_V2_L1

Corey performed an experiment about the water temperature. He mixed different amounts of the chemical with water in a test tube and measured the temperature of the water. The results of his experiment are shown in the table below.

Trail	Amount of Water	Amount of Chemical	Temperature of Water before Adding Chemical	Temperature of Water 2 Minutes after Adding Chemical
1	100 mL	0 grams	21 ºC	21 ºC
2	100 mL	5 grams	21 ºC	27 ºC
3	100 mL	10 grams	21 ºC	32 ºC
4	100 mL	15 grams	21 ºC	35 ºC
5	100 mL	20 grams	21 ºC	35 ºC

4	100 mL	15 grams	21 °C	35 °C
5	100 mL	20 grams	21 °C	35 °C

Which of these is a valid conclusion based on the results of Corey's experiment?

A. Adding more of the chemical will always heat the water to a greater temperature.
B. The chemical always heats water to the same temperature.
C. The temperature of water is not affected by the amount of the chemical.
D. Adding more of the chemical will heat the water but only up to a certain temperature.

Item source: Louisiana YR2005 Gr8

5	100 mL	20 grams	21 °C	35 °C

Which of these is a valid conclusion based on the results of Corey's experiment?

A. The chemical always heats water to the same temperature.
B. The temperature of water is not affected by the amount of the chemical.
C. Adding more of the chemical will always heat the water to a greater temperature.
D. Adding more of the chemical will heat the water but only up to a certain temperature.

Which of these is a valid conclusion based on the results of Corey's experiment?

A. The chemical always heats water to the same temperature.
B. The temperature of water is not affected by the amount of the chemical.
C. Adding more of the chemical will always heat the water to a greater temperature.
D. Adding more of the chemical will heat the water but only up to a certain temperature.

LS_D2_H

Andrea is investigating the effects of light intensity and carbon dioxide concentration on the rate of photosynthesis. She measured the rate of photosynthesis at different light intensities for two identical plants. The plants were placed in closed containers.

LS_D2_V1_L2

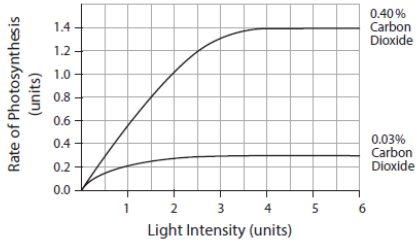
Andrea is investigating the effects of light intensity and carbon dioxide concentration on the rate of photosynthesis.

She measured the rate of photosynthesis at different light intensities for two identical plants. The plants were placed in closed

LS_D2_V2_L2

Andrea is studying photosynthesis. She measured the rate of photosynthesis at different light intensities for two identical plants. The plants were placed in closed containers. One container had an initial carbon dioxide concentration of 0.40%. The other container had an initial carbon

One container had an initial carbon dioxide concentration of 0.40%. The other container had an initial carbon dioxide concentration of 0.03%. She plotted her results as shown below.



Look at the graph.

A. Does an increase in carbon dioxide concentration affect the rate of photosynthesis? (Check one box.)

Yes

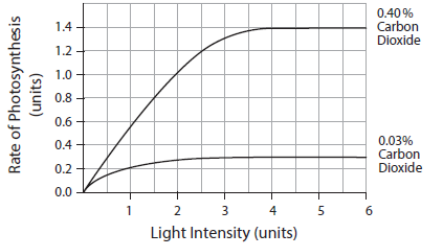
No

B. Explain your answer.

Item source: TIMSS YR2011 Gr8 S042022

containers. One container had an initial carbon dioxide concentration of 0.40%. The other container had an initial carbon dioxide concentration of 0.03%.

She plotted her results as shown below.

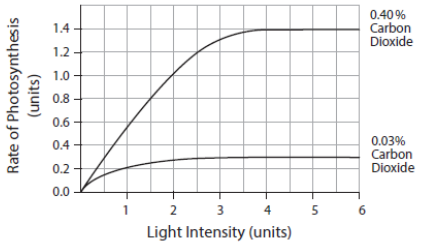


Look at the graph, which of the following statement is correct?

A. The more light the faster the rate of photosynthesis.
B. Carbon dioxide levels may rise to their initial height, but they don't seem to affect the unit of photosynthesis as shown in the graph.
C. Carbon dioxide is not necessary for photosynthesis. An increase in carbon dioxide concentration doesn’t affect the rate of photosynthesis.

dioxide concentration of 0.03%.

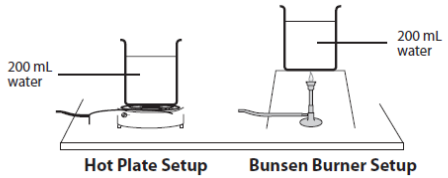
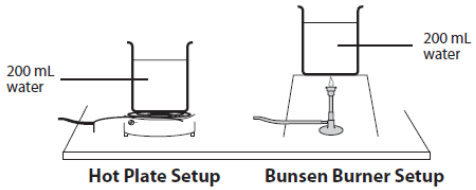
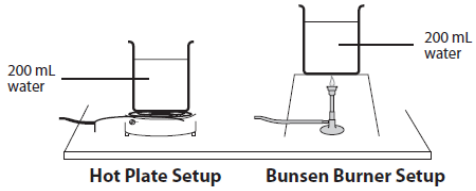
She plotted her results as shown below.



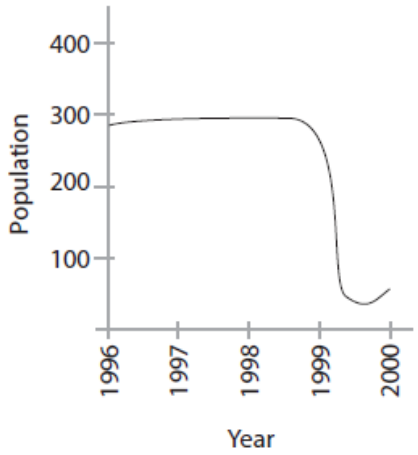
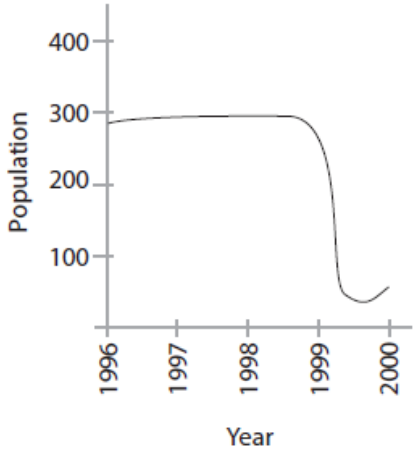
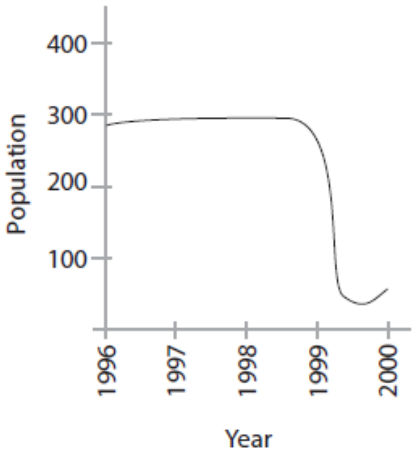
Look at the graph, which of the following statement is correct?

A. The more light the faster the rate of photosynthesis.
B. Carbon dioxide levels may rise to their initial height, but they don't seem to affect the unit of photosynthesis as shown in the graph.
C. Carbon dioxide is not necessary for photosynthesis. An increase in carbon dioxide concentration doesn’t affect the rate of photosynthesis.
D. Carbon dioxide is required for photosynthesis. The higher the concentration of carbon dioxide the faster the rate of

	D. Carbon dioxide is required for photosynthesis. The higher the concentration of carbon dioxide the faster the rate of photosynthesis.	photosynthesis.
PS_D2_L* You want to cook an egg in boiling water at the top of a mountain that is much higher than where you normally live. What, if anything, would you change in your normal cooking procedures in order to cook the egg to the same hardness? Explain why. Item source: NAEP YR2005 Gr12, 2005-12S13 #6	PS_D2_V1_L1 Jason wants to study the relation of boiling temperature to air pressure or altitude. He cooks some pasta in boiling water at the top of a mountain. The top of the mountain is much higher than where he normally lives. Compared to his normal cooking procedure, Jason has to cook the pasta for a longer time before it is ready to eat. Which statement is correct? A. It takes longer to boil water at high altitudes, so it takes a longer time to cook the pasta until it is ready to eat. B. Water boils at a lower temperature than normal because the air pressure at high altitudes is much lower than normal. C. There isn't enough heat to cook the pasta until it is ready to eat due to the lack of enough oxygen at high altitudes to set up the fire.	PS_D2_V2_L1 Jason cooks some pasta in boiling water about 10 minutes at where he normally lives. He cooks it in boiling water about 20 minutes at the top of a mountain. Compared to his normal cooking procedure, Jason has to cook the pasta for a longer time before it is ready to eat. Which statement is correct? A. It takes longer to boil water at high altitudes, so it takes a longer time to cook the pasta until it is ready to eat. B. Water boils at a lower temperature than normal because the air pressure at high altitudes is much lower than normal. C. There isn't enough heat to cook the pasta until it is ready to eat due to the lack of enough oxygen at high altitudes to set up the fire. D. The mountain is much closer to the sun compared with the normal cooking place, so it takes

	D. The mountain is much closer to the sun compared with the normal cooking place, so it takes longer time to cook the pasta until it is ready to eat.	longer time to cook the pasta until it is ready to eat.
PS_D2_H Two kinds of heat sources are usually available in the science lab; an electric hot plate and a Bunsen burner. Jack planned an investigation to test which of these sources heats water faster. He poured 200 mL of water into each of two identical beakers and recorded the initial temperature of the water in each beaker. Jack then placed one beaker on a hot plate and the other over a Bunsen burner, as shown below.  He recorded the temperature of the water in each set up every two minutes for ten minutes. List one variable that Jack controlled in his investigation.	PS_D2_V1_L2 Two kinds of heat sources are usually available in the science lab; an electric hot plate and a Bunsen burner. Jack planned an investigation to test which of these sources heats water faster. He poured 200 mL of water into each of two identical beakers and recorded the initial temperature of the water in each beaker. Jack then placed one beaker on a hot plate and the other over a Bunsen burner, as shown below. He recorded the temperature of the water in each set up every two minutes for ten minutes.  What variables did Jack control in his investigation?	PS_D2_V2_L2 Two kinds of heat sources are usually available in the science lab; an electric hot plate and a Bunsen burner. Jack poured 200 mL of water into each of two identical beakers and recorded the initial temperature of the water in each beaker. Jack then placed one beaker on a hot plate and the other over a Bunsen burner, as shown below. He recorded the temperature of the water in each set up every two minutes for ten minutes.  What variables did Jack control in his investigation? A. Beakers and the initial temperature of the water.

Item source: TIMSS YR2011 Gr8 S042238B	A. Beakers and the initial temperature of the water. B. Timing and the temperature of heat sources. C. Beakers and the amount of water. D. The temperature of heat sources and the location of the experiment.	B. Timing and the temperature of heat sources. C. Beakers and the amount of water. D. The temperature of heat sources and the location of the experiment.

LS_D3_L	LS_D3_V1_L1	LS_D3_V2_L1
		
The graph indicates the number of antelopes in a certain area over a period of time. Which of the following factors is most likely to have caused the sudden change in population between 1999 and 2000? A. global warming B. absence of predators C. depletion of the ozone layer D. brush fires that destroyed the food supply Item source: TIMSS YR2011 Gr8 S032315	In an ecosystem, living organisms and nonliving components of their environment interact as a system. The graph above indicates the number of antelopes in a certain area over a period of time. Which of the following factors is most likely to have caused the sudden change in population between 1999 and 2000? A. global warming B. absence of predators C. depletion of the ozone layer D. brush fires that destroyed the food supply	The graph above indicates the number of antelopes in a certain area over a period of time. Which of the following factors is most likely to have caused the sudden change in population between 1999 and 2000? A. global warming B. absence of predators C. depletion of the ozone layer D. brush fires that destroyed the food supply

LS_D3_H

An ecosystem that retains a high biodiversity (that is, a wide variety of living things) is much more likely to adapt to human-caused environment change than is one that has little. Consider the two food webs shown in the diagram. The arrows point from the organism that gets eaten to the one that eats it. These food webs are highly simplified compared with food webs in real ecosystems, but they still illustrate a key difference between more diverse and less diverse ecosystems.

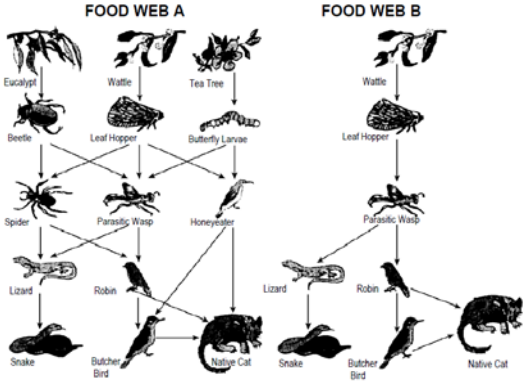
Food web B represents a situation with very low biodiversity, where at some levels the food path involves only a single type of organism. Food web A represents a more diverse ecosystem with, as a result, many more alternative feeding pathways.

Generally, loss of biodiversity should be regarded seriously, not only because the organisms that have become extinct represent a big loss for both ethical and utilitarian (useful benefit) reasons, but also because the organisms that remain have become more vulnerable (exposed)

LS_D3_V1_L2

An ecosystem that retains a high biodiversity (that is, a wide variety of living things) is much more likely to adapt to human-caused environment change than is one that has little. Consider the two food webs shown in the diagram.

Food web B represents a situation with very low biodiversity, where at some levels the food path involves only a single type of organism. Food web A represents a more diverse ecosystem with, as a result, many more alternative feeding pathways.

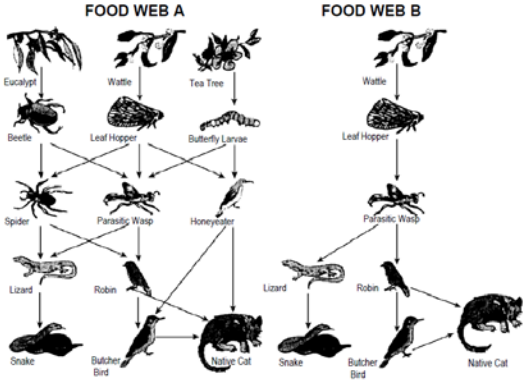


Food webs A and B are in different locations. Imagine if Leaf Hoppers died out in both locations. Which one of these is the **best** prediction and explanation for

LS_D3_V2_L2

Consider the two food webs shown in the diagram.

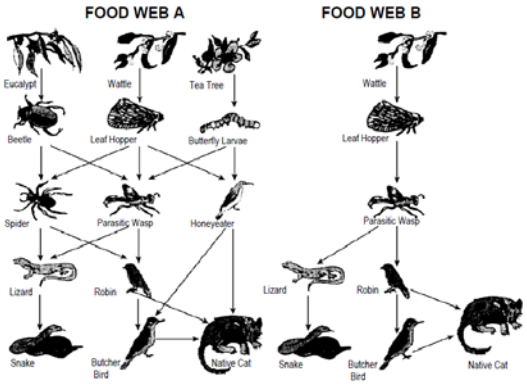
Food web B represents a situation with very low biodiversity, where at some levels the food path involves only a single type of organism. Food web A represents a more diverse ecosystem with, as a result, many more alternative feeding pathways.



Food webs A and B are in different locations. Imagine if Leaf Hoppers died out in both locations. Which one of these is the **best** prediction and explanation for the effect this would have on the food webs?

A. The effect would be greater in

to extinction in the future.



Food webs A and B are in different locations. Imagine if Leaf Hoppers died out in both locations. Which one of these is the best prediction and explanation for the effect this would have on the food webs?

A. The effect would be greater in food web A because the Parasitic Wasp has only one food source in web A.
B. The effect would be greater in food web A because the Parasitic Wasp has several food sources in web A.
C. The effect would be greater in food web B because the Parasitic Wasp has only one food source in web B.

the effect this would have on the food webs?

A. The effect would be greater in food web A because the Parasitic Wasp has only one food source in web A.
B. The effect would be greater in food web A because the Parasitic Wasp has several food sources in web A.
C. The effect would be greater in food web B because the Parasitic Wasp has only one food source in web B.
D. The effect would be greater in food web B because the Parasitic Wasp has several food sources in web B.

food web A because the Parasitic Wasp has only one food source in web A.
B. The effect would be greater in food web A because the Parasitic Wasp has several food sources in web A.
C. The effect would be greater in food web B because the Parasitic Wasp has only one food source in web B.
D. The effect would be greater in food web B because the Parasitic Wasp has several food sources in web B.

D. The effect would be greater in food web B because the Parasitic Wasp has several food sources in web B. Item source: PISA YR2000 Gr8 S126Q04		
PS_D3_L The temperature in the Grand Canyon ranges from below 0 ° C to over 40 ° C. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water in rock cracks help to speed up the breakdown of rocks? A. Freezing water dissolves warm rocks. B. Water cements rocks together. C. Ice smoothes the surface of rocks. D. Freezing water expands in the rock cracks. Item source: PISA YR2006 Gr8	PS_D3_V1_L1 The temperature in the Grand Canyon ranges from below 0 ºC to over 40 ºC. Temperature changes the state of water. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water in rock cracks help to speed up the breakdown of rocks? A. Freezing water dissolves warm rocks. B. Freezing water expands in the rock cracks. C. Ice smoothes the surface of rocks. D. Water cements rocks together.	PS_D3_V2_L1 The temperature in the Grand Canyon ranges from below 0 ºC to over 40 ºC. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water in rock cracks help to speed up the breakdown of rocks? A. Freezing water dissolves warm rocks. B. Freezing water expands in the rock cracks. C. Ice smoothes the surface of rocks. D. Water cements rocks together.

PS_D3_H	PS_D3_V1_L2	PS_D3_V2_L2
For drinks during the day, Peter has a cup of hot coffee, at a temperature of about 90 ºC, and a cup of cold mineral water, with a temperature of about 5 ºC. The cups are of identical type and size and the volume of each drink is the same. Peter leaves the cups sitting in a room where the temperature is about 20 ºC. What are the temperatures of the coffee and the mineral water likely to be after 10 minutes? A. 70 ºC and 10 ºC B. 90 ºC and 5 ºC C. 70 ºC and 25 ºC D. 20 ºC and 20 ºC Item source: PISA YR2006 Gr8	Heat energy moves from warmer objects to colder ones. Peter has a full cup of hot coffee, at a temperature of about 90 ºC, and a full cup of cold mineral water, with a temperature of about 5 ºC. The cups are of identical type and size and the volume of each drink is the same. Peter leaves the cups sitting in a room where the temperature is about 20 ºC. What are the temperatures of the coffee and the mineral water likely to be after 10 minutes? A. 70 ºC and 10 ºC B. 90 ºC and 5 ºC C. 70 ºC and 25 ºC D. 20 ºC and 20 ºC	Peter has a full cup of hot coffee, at a temperature of about 90 ºC, and a full cup of cold mineral water, with a temperature of about 5 ºC. The cups are of identical type and size and the volume of each drink is the same. Peter leaves the cups sitting in a room where the temperature is about 20 ºC. What are the temperatures of the coffee and the mineral water likely to be after 10 minutes? A. 70 ºC and 10 ºC B. 90 ºC and 5 ºC C. 70 ºC and 25 ºC D. 20 ºC and 20 ºC

Note. * Minor mistakes were made for manipulations in this pair (PS_D2_V1_L1, PS_D2_V2_L1).

Appendix B: Item Parameter *b* T-test Results

Item ID	*b*	s.e.	Confidence Interval	Item ID	*b*	s.e.	Confidence Interval	T-test results
ESLs								
LS_D1_V1_L	-0.92	0.18	(-1.39, -0.45)	LS_D1_V2_L	-1.64	0.19	(-2.15, -1.14)	Not significant
PS_D1_V1_L	-0.27	0.17	(-0.73, 0.18)	PS_D1_V2_L	-1.66	0.19	(-2.17, -1.16)	significant
LS_D1_V1_H	1.24	0.20	(0.72, 1.76)	LS_D1_V2_H	2.08	0.21	(1.53, 2.63)	Not significant
PS_D1_V1_H	1.45	0.20	(0.91, 1.99)	PS_D1_V2_H	0.67	0.16	(0.25, 1.10)	Not significant
LS_D2_V1_L	-2.44	0.23	(-3.05, -1.84)	LS_D2_V2_L	-1.20	0.19	(-1.69, -0.70)	significant

PS_D2_V1_L	-1.76	0.19	(-2.28, -1.25)	PS_D2_V2_L	-1.05	0.18	(-1.53, -0.57)	Not significant
LS_D2_V1_H	-0.34	0.16	(-0.75, 0.08)	LS_D2_V2_H	0.22	0.17	(-0.24, 0.67)	Not significant
PS_D2_V1_H	0.40	0.16	(-0.05, 0.79)	PS_D2_V2_H	0.89	0.18	(0.40, 1.38)	Not significant
LS_D3_V1_L	1.26	0.18	(0.79, 1.72)	LS_D3_V2_L	0.97	0.19	(0.48, 1.47)	Not significant
PS_D3_V1_L	-0.28	0.17	(-0.73, 0.18)	PS_D3_V2_L	-0.73	0.16	(-1.16, -0.30)	Not significant
LS_D3_V1_H	-0.21	0.16	(-0.63, 0.20)	LS_D3_V2_H	0.44	0.17	(-0.03, 0.90)	Not significant