Assessing the Development of Students' Statistical Thinking: An Exploratory Study

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Laura Le

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Joan Garfield, Ph.D., Advisor
Andrew Zieffler, Ph.D., Co-Advisor

January 2017

Acknowledgements

I would not be where I am without the guidance, support, encouragement, and friendship from my two advisors, Dr. Joan Garfield and Dr. Andy Zieffler. I started as a student seeking knowledge on how to better teach statistics and leave as a student seeking knowledge to unanswered statistics education questions. You both have molded my thinking AND especially my teaching. I hope to light the fire in other aspiring statistics educators as you have done for me.

To my husband, thank you for putting up with me as a never-ending student all these years. I am deeply grateful for your patience and support and for encouraging me pursue my passion. Also, thank you to my mom and dad! From a young age, you promoted my love of learning and taught me about how to be kind to others. I carry both of those with me everyday in everything I do. And mom, you were right when you said that I should be a teacher. I just was not ready to listen to your words of wisdom when I was in high school and college.

I am also thankful for all of my friends and colleagues. Thank you for listening to me when times were stressful, for providing me feedback on my drafts and presentations, and for building up my confidence on completing this big milestone. Especially thanks to Ann Brearley, my career mom, for keeping tabs on me; to Eric Weber, for providing wise words of wisdom; and to my duplicate self, Laura Ziegler, for being there for brainstorming times and keeping me sane! Also, a big thanks to all of my statistics education family! There are not words for how you all have helped me grow as a student and researcher of statistics education.

Abstract

Developing students' statistical thinking has been stressed as an important learning objective for statistics courses. In general, statistical thinking has been defined as "thinking like an expert applied statistician." However, there is currently no consensus on the characteristics that make up statistical thinking. In addition, there is no known assessment that measures the complete construct of statistical thinking.

The purpose of this study was to assess students' statistical thinking in an introductory statistics course that is based on modeling and simulation. Specifically, the research question of interest was *what components of students' statistical thinking are revealed and developed in an introductory course that is based on modeling and simulation?* To assess this, an assessment was created, called Modeling to Elicit Statistical Thinking (MODEST), that was based on a model of statistical thinking and utilized a type of problem that has been suggested to assess expert-like thinking (i.e., a Model-Eliciting Activity; MEA). To try to ensure that MODEST was an assessment of statistical thinking, several phases of feedback and pilot testing were carried out during the assessment development phase.

In the field test phase, MODEST was administered online twice, at the beginning and at the end of the semester, to students enrolled in an introductory course that is based on modeling and simulation. Responses from 88 students were scored using a detailed scoring rubric to answer the research question. The results indicated that students appeared to enter the course with a moderate amount of statistical thinking (average score = 52%) and leave having developed some statistical thinking as a result of the course

(average score difference = 6%; 95% CI: 2% to 10%). Even though the increase in their overall statistical thinking was significant, it was moderate (Cohen's $d = 0.34$). Based on this, it appears that more could be done in the course to increase students' statistical thinking.

MODEST can be a valuable addition to the statistics education community by filling in the gap of assessing students' statistical thinking. Both statistics education researchers and instructors would benefit from using MODEST to understand statistical thinking.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The central point of education has been described as teaching people to think, to use their rational powers, and to become better problem solvers (Gagné, 1980). It can also be argued that developing higher-order thinking skills in students is a goal for education. These skills include "critical, logical, reflective, metacognitive, and creative thinking" (King, Goodson, Rohani, n.p.).

In the field of statistics, the importance of developing students' statistical thinking–a type of higher-order thinking construct–has been stressed. This recommendation has been made by statisticians and statistics educators (e.g., Moore, 1997; Snee, 1993; Wild, 1994) as well as influential education documents (e.g., Guidelines for Assessment and Instruction in Statistics Education College (*GAISE*) report, American Statistical Association, 2016; Common Core Standards, National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). These recommendations are a part of a larger teaching reform movement in the mathematical sciences, which call for a change in pedagogy and content to better prepare students for the real world (Moore, 1997; American Statistical Association, 2016).

Although the recommendation of teaching statistical thinking applies to statistics courses at all levels (i.e., K-12 to college), changes in introductory statistics courses at the undergraduate level have received particular attention, in part because of the increase in the number of students that enroll in these courses each year (see American Statistical

Association, 2016). Another reason for changes at this level is due to the changes that have been made at the secondary-level of education. More students are exposed to statistics in the classroom at earlier ages and, as a result, are bringing in more prior knowledge than previous generations. In addition, changes have also occurred because of the need to change the curriculum to better help students focus on the "big ideas" of statistics (e.g., deal with variability, understand the logic of inference) and not teach statistics as a cookbook approach, such as teaching statistical methods as a set of computations and procedures (e.g., Cobb, 2007; American Statistical Association, 2016).

One major curricular change that is being implemented in introductory statistics courses is a shift in content away from only teaching normal-based methods (e.g., t-tests) toward the use of modeling and simulation to teach inference. Several curricula have been created using modeling and simulation, including the CATALST course (Garfield, delMas, & Zieffler, 2012), the Lock textbook (Lock, Lock, Lock Morgan, Lock, & Lock, 2013), and the ISI textbook (Tintle et al., 2013). Arguments for implementing this change include focusing on the logic of inference via simulation rather than normal approximations to sampling distributions (Cobb, 2007) and promoting ways of thinking statistically (Garfield, et al., 2012).

1.1 Understanding Statistical Thinking

Even with the emphasis on teaching students to think statistically, there is currently no consensus on the characteristics that make up statistical thinking. In general, statistical thinking has been described as "thinking like an expert applied statistician." To try to further clarify what statistical thinking means, researchers have attempted to define

statistical thinking based on their own opinion. For example, Snee (1993) described statistical thinking as understanding that systems are interconnected, recognizing that variation is everywhere, acknowledging the need for data, and knowing how to use statistical methods to understand variation. Similarly, Moore (1990) wrote that statistical thinking is recognizing the need for data, the design of data collection, the omnipresence of variation, and the quantification and explanation of variation. In contrast, Wild and Pfannkuch (1999) used empirical data to create a four-dimensional framework of statistical thinking, which included the dimensions of investigative cycle, types of thinking, interrogative cycle, and dispositions. Wild and Pfannkuch's research suggests that statistical thinking is more a complex process than a list of four or five broad characteristics.

Understanding how experts think is not unique to the field of statistics. Other domain-specific areas, such as mathematics, and domain-general areas, such as research comparing experts to novices, have investigated what it means to think like an expert. Therefore, to further understand the construct of thinking like an expert, there is a need to examine literature from these areas to understand more broadly how experts think. The characteristics of expert thinking from these other fields could be integrated with Wild and Pfannkuch's framework to provide a more complete model of statistical thinking.

1.2 The Need for an Assessment of Statistical Thinking

The uses of developing a more complete model of statistical thinking are numerous. The most important use would be assessing students' statistical thinking in a classroom that has a learning objective of "develop students' statistical thinking".

Currently, however, there is no known assessment that measures the complete construct of statistical thinking. As a consequence, statistics courses with the learning objective of "develop students' statistical thinking" are not able to assess whether they have met their goal or not. To address this problem, that is, to understand whether students' statistical thinking develop in a statistics course, a quality assessment is needed that measures students' statistical thinking in a course. This assessment should be based off of the more complete construct of statistical thinking that was mentioned previously.

1.3 Description of the Study

This study aimed to answer the following research question:

What components of students' statistical thinking are revealed and developed in an introductory statistics course that is based on modeling and simulation?

An assessment of statistical thinking, called Modeling to Elicit Statistical Thinking (MODEST), was created to assess students' statistical thinking in an introductory statistics course that is based on modeling and simulation. MODEST was based on a model of statistical thinking and utilized a type of problem that has been suggested to assess expert-like thinking (i.e., a Model-Eliciting Activity; MEA). To answer the research question, responses from students in an introductory statistics course that is based on modeling and simulation (i.e., the CATALST course) were used because the course has an explicit learning objective of "develop students' statistical thinking."

1.4 Structure of the Dissertation

Chapter 2 provides a review of the literature related to understanding, developing, and assessing statistical thinking. To understand statistical thinking, literature related to

defining the characteristics of expert thinking in domain-specific areas (i.e., statistics and mathematics) and in domain-general areas are examined. Then, literature related to developing and assessing expert thinking, with a focus on using ill-structured problems, is summarized, followed by a review of the literature related to assessing statistical thinking. Chapter 3 describes the methodology for this study. This includes the development, refinement, and administration of MODEST and its test blueprint. This chapter also reports the data collection and the data analysis for this study.

Chapter 4 presents the results of the development and administration of MODEST. First, three sets of feedback on the assessment are described and the resulting changes to the assessment are summarized. Then, the pilot test administrations are presented and changes to the assessment are reported. Finally, the results of the field test administration are reported using descriptive and inferential methods.

Chapter 5 summarizes the results of the study. In addition, validity evidence for MODEST as an assessment of statistical thinking is presented, as well as limitations of this study and implications for teaching and research.

Chapter 2

Literature Review

2.1 Introduction

The term statistical thinking has generally been considered as *thinking like an expert statistician* in the literature. In order for the development of students' statistical thinking to occur within a statistics course, there is a need to understand the construct of statistical thinking, how statistical thinking can be developed within a statistics course, and how statistical thinking can be assessed within a statistics course. To this end, this review examines relevant literature related to understanding, developing, and assessing statistical thinking. Literature related to understanding statistical thinking is presented first. Then, literature related to developing and assessing expert thinking, with a focus on using ill-structured problems, is summarized. Lastly, literature specific to assessing statistical thinking is presented.

2.2 Literature That Contributes to Understanding Statistical Thinking

To understand how statistical thinking has been described in the literature, descriptions of statistical thinking from three leading perspectives are presented and compared. Then, to further understand how other areas have characterized thinking like an expert, literature related to expert thinking in a domain-specific area (i.e., mathematics) and domain-general areas are also examined.

2.2.1 Defining statistical thinking. Currently, there is no consensus on what statistical thinking means. As a consequence, the term *statistical thinking* is often used interchangeably with two other terms: *statistical literacy* (e.g., Watson, 1997; Ziegler,

2014) and *statistical reasoning* (e.g., Sedlmeier, 2000; Jones et al., 2001; Ben-Zvi & Friedlander, 1997; Chick & Watson, 2002). To clarify the differences among the three terms, authors have attempted to provide definitions of the three (e.g., Ben-Zvi & Garfield, 2004; Chance, 2002). Ben-Zvi and Garfield (2004) defined *statistical literacy* as "basic and important skills that may be used in understanding statistical information or research results…[including] understanding of concepts, vocabulary, and symbols, and…an understanding of probability as a measure of uncertainty" (p. 7). In contrast, they defined *statistical reasoning* as "the way people reason with statistical ideas and make sense of statistical information…reasoning means understanding and being able to explain statistical processes and being able to fully interpret statistical results" (p. 7). Finally, they defined *statistical thinking* as thinking like an applied statistician. This includes understanding the big ideas of statistics (e.g., omnipresence of variability; inference from a sample to a population), using appropriate statistical methods and models, and understanding and carrying out the process of statistical investigations. Chance (2002) expanded on statistical thinking as the "ability to see process as a whole (with iteration)…able to move beyond what is taught in the course, to spontaneously question and investigate the issues and data involved in a specific context" (Definitions of Statistical Thinking section, para. 17).

While the descriptions of statistical thinking offer a general sense of what it means to think statistically, an operational definition of statistical thinking is needed to fully understand this construct. To this end, three leading perspectives that define the components of statistical thinking are presented: Ronald Snee (1990, 1993), David Moore

(1990), and Chris Wild and Maxine Pfannkuch (1999). Two of the three perspectives, Snee and Moore, are based on opinions from prominent statisticians, while the third perspective, Wild and Pfannkuch, is based on the only known empirical study of statistical thinking.

2.2.1.1 Perspective of Ronald Snee. Ronald Snee has been the predominant spokesman in the area of total quality control and has urged for the development of statistical thinking in statisticians, quality professionals, and business leaders. His perspective of statistical thinking has influenced the teaching of statistical methods in Six Sigma training (Hoerl & Snee, 2002; Hoerl, 2001; Hoerl & Snee, 2010; Antony, 2004) and the definition of statistical thinking in a statistical glossary published by the American Society of Quality Statistics Division (American Society of Quality, 1996).

Snee (1990, 1993) characterized statistical thinking as the thought processes that occur when solving problems, improving systems, and predicting future performance on a system. He proposed that statistical thinking is made up of four elements:

- understanding that all work is made up of interconnected processes,
- recognizing that variation occurs in all processes,
- understanding that data are needed to measure variation, and
- knowing how to use statistical methods and tools to identify, quantify, control, reduce, and understand the variation and make predictions.

Furthermore, he believed that the core of statistical thinking is the collection and analysis of data (Snee, 1993).

Snee (1993) also stressed the need to change the content and the delivery of statistics courses to incorporate statistical thinking within the classroom. Content, he remarked, should "place greater emphasis on data collection, understanding and modeling variation, graphical display of data, design of experiments, surveys, problem-solving, and process improvement" (p. 151). He proposed that the delivery of statistical material should be based on an experiential learning approach, which includes choosing contexts of personal interest to the students. Snee believed that changing both the content and delivery would help students experience statistical thinking in real-world situations and develop an appreciation for statistical techniques. Additionally, he posited that these changes would elicit more positive student attitudes, and more importantly, increase the likelihood that students would actually think statistically.

2.2.1.2 Perspective of David Moore. David Moore is a well-known statistician and author of leading introductory statistics textbooks. In the statistics education literature, he has been a prominent contributor on the topic of statistical thinking and has written what is probably the most cited characterization of statistical thinking. His conception of statistical thinking was the basis for the recommendation on statistical thinking in two influential reports in statistics education: the Cobb Report (Cobb, 1992) and the GAISE College Report (American Statistical Association, 2016).

Moore (1990) defined statistical thinking as a general way of thinking in the realm of inquiry. He categorized statistical thinking into five core elements:

- the need for data about processes,
- the design of data production with variation in mind,

- the omnipresence of variation in processes,
- the quantification of variation, and
- the explanation of variation.

Similar to Snee, Moore also argued for statistical thinking to be developed within the statistics classroom. He claimed that statistical thinking is an "independent and fundamental intellectual method" (p. 136), and "will not be developed in children if it is not a present in the curriculum" (p. 135). Moore recommended that students be explicitly taught to deal with variation and uncertainty in data. He also believed that the mental habits of every educated citizen should include statistical thinking.

2.2.1.3 Research by Chris Wild and Maxine Pfannkuch. In contrast to the opinions of statistical thinking presented by Snee and Moore, Chris Wild and Maxine Pfannkuch (1999) conducted an empirical study to investigate the statistical thinking that occurs in applied statisticians. The study consisted of interviewing six applied statisticians from a variety of fields and 16 advanced students who were involved in statistical tasks. The participants were asked to describe their approach and their thinking when solving statistical problems. Based on the common characteristics among the participants' responses, Wild and Pfannkuch developed a theory of the construct of statistical thinking: a four-dimensional framework for statistical thinking in empirical enquiry (see Figure 1). They hypothesized that individuals work in all four dimensions simultaneously when they are involved in tasks that are statistical.

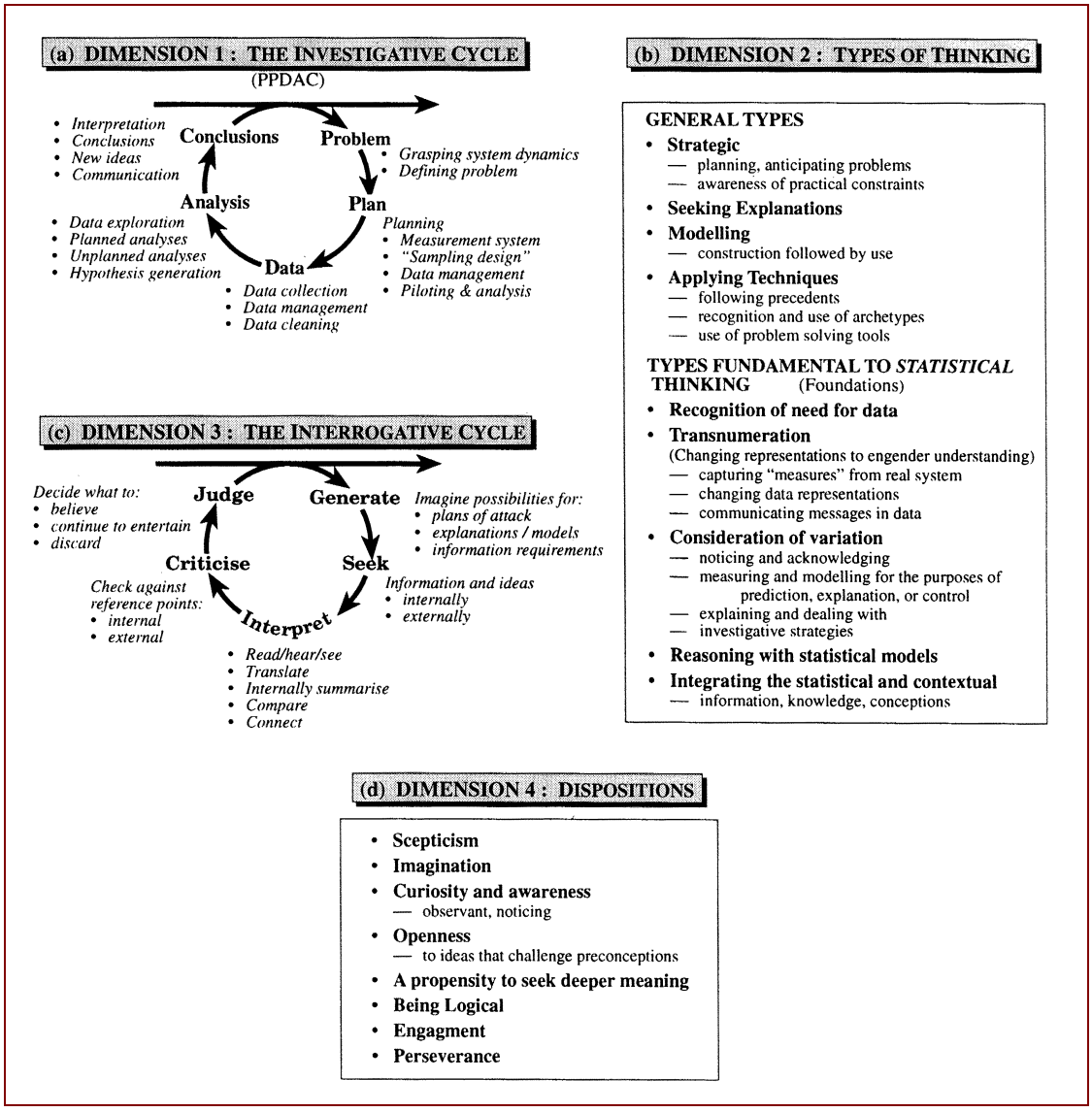

Figure 1. Wild and Pfannkuch's four-dimensional framework for statistical thinking in empirical research. Adapted from "Statistical Thinking in Empirical Enquiry," by C. J. and M. Pfannkuch, 1999, *International Statistical Review*, *67*, p. 226. Copyright 1999 by the International Statistical Institute.

2.2.1.3.1 Framework of statistical thinking. The four dimensions of Wild and Pfannkuch's framework are investigative cycle, types of thinking, interrogative cycle, and dispositions. Each of these dimensions is elaborated on next.

Wild and Pfannkuch (1999) described the investigative cycle dimension as how one acts and thinks when involved in solving a statistical problem. They suggested that five stages make up the investigative cycle: Problem, Plan, Data, Analysis, and Conclusions (PPDAC). This cycle was an adaptation of an earlier model by MacKay and Oldford (1994). Wild and Pfannkuch also claimed that the " PPDAC cycle is concerned with abstracting and solving a statistical problem grounded in a larger "real" problem" (p. 225).

For the dimension of types of thinking, Wild and Pfannkuch (1999) characterized it into two categories. The categories were labeled thinking that was common to all problem-solvers and thinking that was fundamentally statistical. These thinking categories were further divided into subcategories.

For the thinking that was common to all problem-solvers, Wild and Pfannkuch identified four subcategories. The first general type of thinking was strategic thinking; that is, thinking related to developing a logistical plan for solving a problem. Examples of strategic thinking included plan of attack, setting deadlines for subtasks, division of labor, and anticipating problems. The second general type, seeking explanations, was described as coming up with several alternative reasons for the cause of the response. The third general type was modeling. This involved constructing mental and statistical models to understand and learn about real-world phenomena. The final type, application of techniques, consisted of applying and adapting problem-solving strategies to a new problem.

For the thinking that was fundamental to statistics, Wild and Pfannkuch (1999) identified five subcategories. They believed that these types of thinking are "the foundations on which statistical thinking rests" (p. 227). The first foundational type of statistical thinking was recognition of the need for data. It was described as the awareness of using data to make decisions rather than using personal experiences and anecdotal evidence. The second foundational type was transnumeration, which was defined as "a dynamic process of changing representations to engender understanding" (p. 227). Consideration of variation was the third foundational type of statistical thinking. This involved understanding that variation is all around and that statistics is needed to understand the variation. A fourth foundational type was reasoning with statistical models; that is, understanding and using models that are unique to the discipline of statistics. The last foundational type of thinking was the integration of statistical and conceptual knowledge. This subcategory was described as continually going back and forth between the contextual sphere and the statistical sphere.

Wild and Pfannkuch (1999) identified the dimension of interrogative cycle as a thinking process that is used while solving a problem. They proposed five components for this cycle:

- generation of ideas and plans of attack,
- seeking information and ideas,
- interpretation of results and information and making connections,
- criticizing information and ideas, and
- forming a judgment on the information, decision, and ideas.

They state that this cycle was circular, as well as recursive. Wild and Pfannkuch reported that "the thinker is always in one of the interrogative states while problem-solving" (p. 231).

The dimension of dispositions dimension was characterized by Wild and Pfannkuch (1999) as personal qualities that affect the thinking of problem-solvers. They presented eight dispositions, which included curiosity and awareness, imagination, skepticism, being logical, openness, engagement, perseverance, and a propensity to seek deeper meaning.

2.2.1.3.2 Responses to the framework of statistical thinking. Responses to Wild and Pfannkuch's framework of statistical thinking from leading statisticians were solicited in the same issue of the journal as Wild and Pfannkuch's paper (Biehler, 1999; Breslow, 1999; Moore, 1999; Smith, 1999; Snee, 1999). These statisticians were individuals who had written about the topic of statistical thinking, including Snee and Moore. The respondents praised Wild and Pfannkuch's efforts at developing a complex framework that is based on data, but criticized the framework in terms of:

- missing elements within a particular dimension, such as Bayesian thinking (Moore, 1999), a disposition on creativity (Smith, 1999), or a "hypothesis specification" in the investigative dimension (Breslow, 1999),
- needing to expand on the framework, such as adding research on expert thinking from Artificial Intelligence and Cognitive Science (Biehler, 1999) or incorporating relevant statistical tools in the framework (Snee, 1999), and

- having a framework that is too complex, especially for it to be useful in teaching students to think statistically (Snee, 1999; Moore, 1999; Smith, 1999; Breslow, 1999; Biehler, 1999).

2.2.1.3.3 Impact of statistical thinking framework on statistics education. Beyond the framework presented in their 1999 article, Wild and Pfannkuch have contributed to the research on statistical thinking in other ways. Pfannkuch and Wild (2000) further examined the data collected from the interviews with the applied statisticians (described earlier). In this analysis, their goal was to find commonalities between applied statistical practice and statistical thinking. They identified these commonalities and noted that the statisticians did not learn all of them in a statistics course, but rather, learned them over their time as an applied statistician. Some of these missing dimensions in statistics courses included "understanding the dynamics of a system, problem formulation, measurement, and nontechnical aspects of the planning of studies" (p. 151). Based on the results identified in the paper, Pfannkuch and Wild hoped to better inform teachers on how to incorporate such thinking and practice skills into the statistics classroom.

Pfannkuch and Wild (2003) also used their statistical thinking framework to identify students' difficulties when learning to think statistically. Some of the student barriers identified included developing dispositions of statistical thinking, connecting stages of the enquiry cycle, integrating the statistical with the contextual, reasoning with models, dealing with variability, and acknowledging the need for data. These results were used to inform the teachers on how their teaching helps students' to think statistically in their classroom.

Additional work examined statistical thinking from a historical perspective (Pfannkuch & Wild, 2004). Pfannkuch and Wild (2004) used their statistical thinking framework to trace back the origins of statistical thinking and identify contributions from a variety of fields (e.g., psychology, epidemiology, quality management) to the current idea of statistical thinking. The historical perspective helped to understand how statistical thinking has evolved over time.

Building on the work by Wild and Pfannkuch (1999), researchers, statisticians, and statistics educators have used the framework of statistical thinking as a model for developing and assessing statistical thinking in students. Chance (2002) used the four dimensions of their framework to provide guidelines for developing students' mental habits that are necessary for statistical thinking. The guidelines included

- start from the beginning,
- understand the statistical process as a whole,
- always be skeptical,
- think about the variables involved,
- always relate the data to the context,
- understand the relevance of statistics, and
- think beyond the textbook.

Melton (2004) investigated developing students' skepticism, a disposition from Wild and Pfannkuch's framework. She used exercises to develop a healthy skepticism in students when looking at the process of data collection. Researchers also used Wild and Pfannkuch's framework to develop teachers' statistical thinking by creating materials and

designing learning experiences around the components of the framework (e.g., Sanchez & Blancarte, 2008; Makar & Confrey, 2002). Finally, their framework has been used as a basis for assessing statistical thinking (e.g., Chance, 2002; Groth, 2005; Makar & Confrey, 2002; Pfannkuch & Horring, 2005; Pfannkuch & Rubick, 2002), which is discussed later in the literature review.

2.2.1.4 Comparison of the three perspectives of statistical thinking. To summarize how statistical thinking has been defined in the literature, the three perspectives of statistical thinking that have been presented are now compared and contrasted. The similarities include

- describing statistical thinking as the thought processes that occur in expert statisticians when solving statistical problems,
- defining common set of elements of statistical thinking, such as recognizing the need for data, acknowledging that variation is at the core of statistical thinking, and using statistical methods to quantify, explain, model, control, and understand variation,
- emphasizing that statistical thinking is a learning goal for statistics courses at all levels, and
- providing suggestions on how to develop statistical thinking within a statistics course, such as providing students with the opportunity to use statistical thinking in real-world situations (Moore, 1990; Snee, 1993) and stressing the use of trigger questions of *Why?* and *How?* within the classroom (Wild & Pfannkuch, 1999).

In contrast, the three perspectives of statistical thinking differed in their target audience. Snee (1990, 1993) wrote with the business and total quality control audience in mind, whereas Moore (1990) and Wild and Pfannkuch (1999) were targeting, more generally, at introductory statistics students.

Another difference is in the elements that describe statistical thinking. Snee (1990, 1993) believed that statistical thinking included recognizing all work as a process. This view seems unique to the field of total quality and was not included in either of the other two perspectives. Wild and Pfannkuch (1999) also had elements foundational to statistical thinking that was not seen in the other two perspectives. These distinctive elements of transnumeration and integration of the statistical and contextual information were identified in the interviews from applied statisticians. Because of the uniqueness of these elements to Wild and Pfannkuch's model, it may be that these elements are identifiable only when empirical data is used to construct the concept of statistical thinking.

A third distinction among the three perspectives was in the complexity of the descriptions of statistical thinking. Wild and Pfannkuch (1999) presented a more complex concept of statistical thinking relative to the other two perspectives. Rather than focusing on only the elements are foundational to thinking statistically as Snee (1990, 1993) and Moore (1990) did, Wild and Pfannkuch proposed that statistical thinking is made up of an interactive, iterative process that is four-dimensional. This difference in complexity may be explained because Wild and Pfannkuch were the only researchers to base their definition on empirical data. Because of this, an argument can be made that statistical

thinking is more complex than only the types of thinking dimension. Therefore, for the remainder of this literature review, Wild and Pfannkuch's framework of statistical thinking is used as the operational definition of statistical thinking.

Recall that statistical thinking has generally been considered as expert thinking that occurs while solving statistical problems. However, understanding what it means to think like an expert has been examined in other fields, in addition to a broader understanding outside of a specific field (i.e., domain-general). One domain-specific area related to statistics is mathematics. To further understand the concept of thinking like an expert, literature related to understanding how experts think in mathematics is examined next.

2.2.2 Characterizing mathematical thinking. In mathematics education, researchers have tried to characterize the way expert mathematicians think. One researcher, Schoenfeld (1992), described mathematical thinking as "having a mathematical point of view–seeing the world in ways like mathematicians do" (p. 19). Part of this requires the thinker to develop the habits and dispositions of experts, in addition to learning the skills, strategies, and knowledge of mathematics. The learning process of thinking like an expert is a socialization process, known as enculturation (Resnick, 1989). Schoenfeld also summarized areas of cognition that he deemed necessary to be classified as thinking mathematically. These areas were

- having a vast amount of content knowledge and knowing how to use that knowledge,

- being able to use different problem solving strategies in a flexible and innovative way,
- having self-monitoring skills,
- having positive beliefs and affects towards mathematics, and
- being in an environment that promotes the skills, strategies, and knowledge of mathematics, in addition to the habits and dispositions of problem solving.

To further describe mathematical thinking, Schoenfeld (1998) compared thinking like a mathematician to thinking like a cook. He wrote that with experience and time, both expert mathematicians and expert cooks exhibit four similar characteristics of thinking:

- ability to adjust their methods to fit the current situation,
- have access to knowledge of a variety of methods and reference examples,
- ability to recognize features of the problem that alerts them for how to proceed, and
- ability to acknowledge conditionalized information for the problem at hand.

Another researcher summarized a variety of viewpoints on mathematical thinking to answer the question "What is mathematical thinking?" (Sternberg, 1996). There appeared to be no single model for understanding mathematical thinking. Consequently, Sternberg proposed that multiple points of view should be accounted for when trying to understand the construct of mathematical thinking. To do this, a triarchic theory of human intelligence was used. The triarchic theory is composed of a componential subtheory (e.g., processes relevant to intelligent thought), an experiential subtheory (e.g.,

dealing with novelty and making their knowledge automatic), and a contextual subtheory (e.g., having the ability to adapt knowledge to a variety of contexts).

Researchers have also tried to characterize mathematical thinking by empirically studying the construct. One area of research involved investigating how expert mathematicians solve problems. Studies in this area either looked at specific heuristics (e.g., Stylianou, 2002) or general problem-solving processes (e.g., Carlson & Bloom, 2005; Misfeldt & Johansen, 2015) that occurred while the experts solved mathematical problems. Using the data from the expert mathematicians and from theories on problem-solving in mathematics, models of mathematical thinking were created. These models attempted to describe the complex multidimensional cognitive and metacognitive processes that occur while expert mathematicians solve mathematical problems.

The other area of research on mathematical thinking consisted of investigating the characteristics of experts and novices while they solved mathematical problems (e.g., Stylianou & Silver, 2004; Schoenfeld & Herrmann, 1982; DeFranco 1996). Experts, in contrast to novices, were found to have a larger knowledge base and better organization of knowledge. For example, Stylianou and Silver (2004) found that experts had a vast amount of knowledge of how to use visual representations as a tool to solve mathematical problems. Additionally, experts were also found to have a deep understanding of the content. Novices tend to categorize problems based on surface structures of the problem (i.e., naïve characterization of problems based on the most prominent mathematical features in the problem or in the general subject area) whereas experts tend to categorize problems based on deep structures (i.e., mathematical principle needed for solution)

(Schoenfeld and Herrmann, 1982). Lastly, Schoenfeld's 1992 theory on the skills needed for mathematical problem-solving expertise was empirically studied (DeFranco, 1996). Based on data from experts and novices, evidence was found that lend strong support for Schoenfeld's theory on mathematical expertise.

To summarize the literature on describing mathematical thinking, there appears to be no consensus on what it means to think mathematically. However, similar to the general definition of statistical thinking, mathematical thinking has generally been described as a type of expert thinking; specifically, expert thinking that occurs while solving mathematical problems. To learn more about how experts think while solving problem in more general areas, literature related to understanding how experts think compared with novices in domain-general areas is now examined.

2.2.3 Understanding expert verses novice thinking. Researchers have studied differences in thinking between novices and experts in general. These studies help to understand how experts think about and solve problems in their field and to provide insight on how expert thinking can be developed in novices. This section focuses on what the research suggests about the thinking between experts and novices in problem-solving situations.

One general distinction between experts and novices is the amount of domain-specific experience and knowledge each has (Schenk, Vitalari, & Davis, 1998). As a result, experts have more confidence and are better able to adapt to new and unexpected situations than novices. This expertise, however, is domain-specific, rather than domain-general (Bedard & Chi, 1992). Voss, Tyler, and Yengo (1983) found that experts

provided similar solutions as novices when they solved problems from a different domain.

2.2.3.1 Ill-structured problems. Experts, including expert statisticians, commonly encounter problems that are complex and ill-structured as opposed to simple and well-structured. Ill-structured problems are characterized as having

- ill-defined problem conditions;
- no direct link to a particular concept, rule, or principle to solve the problem;
- an interpersonal component to solving the problem due to the need of using personal opinion or beliefs to solve the problem;
- multiple solutions without a prescribed solution path or no solution at all; and
- the problem-solvers make judgments about the problem and justify their solution (Jonassen, 1997).

In contrast, well-structured problems are characterized as having (a) well-defined problem conditions; (b) a finite number of concepts, rules, and principles that can be applied to the situation; and (c) one correct solution, with often a preferred, prescribed solution process (Jonassen, 1997). Well-structured problems are commonly encountered in school and university environments, whereas, ill-structured problems are more routinely encountered in everyday and professional contexts.

Well-structured and ill-structured problems are also different with respect to the transferability of skills needed to solve the problem. There is an assumption that the skills developed by solving well-structured problems transfer to those used in solving ill-structured problems. However, this assumption has been argued to be unreasonable

(Jonassen, 1997; Kitchener, 1983; Schraw, Dunkle, & Bendixen, 1995; Shin, Jonassen, & McGee, 2003; Chi & Glaser, 1985; Sinnott, 1989; Voss & Post, 1988; Voss, Wolfe, Lawrence, & Engle, 1991; Woods et al., 1997). As an example, it has been found that the skills needed to solve well-structured problems are often domain-specific (Glaser & Chi, 1988), whereas the skills needed to solve ill-structured problems tend to be far less domain-specific (Kramer & Woodruff, 1986; Kuhn, 1991; Perkins, Faraday, & Bushey, 1991). That is, the skills from ill-structured problems can be used in problems across different domains.

While the characteristics provide useful guidance for categorizing problems, well-structured and ill-structured problems should not be thought of as dichotomous entities. Instead, problems should be considered as lying on a spectrum, likely sharing characteristics of both types (Reitman, 1965). The degree to which a problem is well- or ill-structured is determined by the combination of four attributes:

- the complexity of the problem,
- the clarity of the goal and the criteria addressing it,
- the number of directions given for the necessary domain skills, and
- the number of possible solutions and/or solution paths.

2.2.3.2 Problem-solving cognitive processes for experts and novices. Differences in cognitive processes have also been found between experts and novices. One cognitive process that differs between experts and novices is the metacognitive skills that they possess (Bransford et al., 2000). Metacognition refers to the ability to monitor and adapt one's decisions while solving a problem. Experts have developed essential metacognitive

skills over time to help them solve problems accurately and with few errors (e.g., Schenk et al., 1998). These skills include questioning and elaborating on their knowledge of the problem, considering counter-examples to help decide whether their knowledge is accurate, and using troubleshooting tactics to correct their knowledge. Novices, on the other hand, tend to lack experience that is needed for metacognitive skills to fully develop in their domain. As a result, they are not able to accurately and effectively monitor their decision-making process and thus are more prone to making common errors while solving a problem.

Another difference between experts and novices is in their organization of domain-specific knowledge. Experts tend to have knowledge structures (e.g., schemas) that reflect a deep understanding of the content (Bransford et al., 2000; Bedard & Chi, 1992). Rather than having their knowledge organized around lists of facts and formulas, experts' knowledge is often structured around the big ideas in their field. Consequently, experts are able to easily navigate in their discipline, to filter irrelevant information from important features in a problem, and to categorize problems around the big ideas in their field (e.g., Schenk et al., 1998; Chi, Feltovich, & Glaser, 1981; Hardiman, Dufresne, & Mestre, 1989).

Experts and novices also differ in their amount of conditionalized knowledge. Conditionalized knowledge is the ability to retrieve the appropriate technique that is relevant to solve a problem (e.g., Bransford et al., 2000, Chi et al., 1981). Relative to novices, this type of knowledge allows experts to retrieve information from memory more automatically and fluently (Bransford et al., 2000). Moreover, when experts are

able to transfer knowledge from one problem to another, they are able to lessen their cognitive demands and at the same time more easily classify and solve problems of similar nature. This "ease of processing" (Bransford et al., 2000, p. 44) is what makes the problem-solver capable of taking in more information about the problem.

2.2.3.3 Problem-solving techniques for experts and novices. Besides cognitive differences between experts and novices, differences have also been observed in the way that novices and experts solve problems. One difference is that experts tend to skip steps or collapse multiple steps into a single step when solving problems (e.g., Blessing & Anderson, 1996; Larkin, McDermott, Simon, & Simon, 1980; Korf, 1985). They also employ a different the type of strategy when solving problems. Two common strategies for problem-solving are the working backward approach and the working forward approach. The working backward approach is described as first selecting a principle without knowing if the given variables connect to the unknown variables (i.e., working from the unknowns to the givens). Conversely, the working forward approach is described as first identifying the known variables and then selecting the principle that uses the known variables to solve for the unknown variables (i.e., working from the givens to the unknowns). Results suggest that novices tend to use the working backward approach to solve problems, while experts tend to use a working forward approach (e.g., Larkin et al., 1980; MacKay & Elam, 1992; Bedard & Chi, 1992). However, one study found that both novices and experts used the working forward approach to solve a problem in physics but each group used different cues from the problem to help solve the problem (Chi, Feltovich, and Glaser, 1981).

As previously mentioned, experts frequently encounter problems that are complex and ill-structured. Because of this, experts and novices differ on how they solve problems. One of the major differences in problem-solving techniques when solving ill-structured problems is problem representation (Bedard & Chi, 1992; Spector, 2006; Ho, 2000; Voss, et al., 1983; Ertmer et al., 2008). Experts tend to try to figure out the goals and givens in the problem statement, come up with challenges and constraints for solving the problem, and break the ill-structure nature of problem into multiple well-structured problems. Novices, on the other hand, try to solve the problem directly without defining it. These differing tactics consequently affect the solutions that are generated and offered by novices and experts (Fernandes & Simon, 1999).

To elaborate on the problem-solving techniques experts use, Jonassen (1997) proposed a seven-step model on how experts solve ill-structured problems (see Table 1). His process was based off an empirical-based model of how experts solve ill-structured problems (Sinnott, 1989). Jonassen further expanded on each step by describing the expert cognitive processes that are needed while solving ill-structured problems.

Table 1

Jonassen's (1997) Seven-Step Problem-Solving Process for Ill-Structured Problems

Problem-Solving Process
1. Articulate problem space and contextual constraints
2. Identify and clarify alternative opinions, positions, and perspectives of stakeholders
3. Generate possible problem solutions
4. Assess the viability of alternative solutions by constructing arguments and articulating personal beliefs
5. Monitor the problem space and solution options
6. Implement and monitor the solution

7. Adapt the solution.

Due to the nature of ill-structured problems, the problem-solver may not have encountered a similar problem in the past. As a result, translating skills and knowledge to the problem at hand can be challenging. However, experts were found to be flexible in their approach to new situations, which is known as adaptive expertise (Bransford et al., 2000). By not following a memorized fixed recipe for solving problems, experts are able to adapt and transfer their knowledge to solve new problems that, as a result, allows them to employ innovative approaches and multiple strategies to solve a problem.

2.2.3.4 Problem-solving by experts and novices in statistics. Only one study on experts and novices was found in the statistics education literature. Alacaci (2004) compared the knowledge of expert statisticians (practicing statisticians and mathematical statisticians) and novice statisticians (doctoral students) in selecting statistical techniques for a variety of research scenarios. Although it was found that there were few differences in knowledge of research design methods, theoretical statistics, and procedural statistics between experts and novices, experts had better connections than novices of statistical knowledge when selecting appropriate statistical techniques for research scenarios and when comparing and contrasting statistical techniques.

2.2.3.5 Summary of expert verses novice thinking in problem-solving situations. Based on the literature on comparing problem-solving between experts and novices, experts are differentiated from novices with respect to the amount of experience and knowledge they have, the cognitive processes they utilize (e.g., metacognition, efficiency,

organization of knowledge, conditionalized knowledge), and problem-solving techniques they employ (e.g., skipping steps, working forward, adaptive expertise). The type of problems encountered by experts and novices tend to be different, with experts commonly encountering ill-structured problems whereas novices commonly encounter well-structured problems. As a possible consequence, experts are different from novices in how they represent problems. Finally, the expert skills needed to solve problems in one domain do not necessarily transfer to other domains.

Based on the literature reviewed, expertise has been identified as a combination of a vast amount of knowledge and experience, well-organized knowledge structures, and expert thinking processes, and expert problem-solving techniques within a field. As a result, these characteristics have helped experts successfully solve novel, often ill-structured, problems. Thus, for expert thinking in problem-solving situations to be developed and assessed, novices should encounter expert-like problems, namely, problems that are complex and ill-structured. The next section examines the literature related to strategies to develop expert thinking using ill-structured problems.

2.3 Strategies to Develop Expert Thinking

Gagné proposed that "the central point of education is to teach people to think, to use their rational powers, to become better problem solvers" (1980, p. 85). Additionally, Bransford (1994) argued that expertise and wisdom cannot be directly taught but must be acquired through experience. Thus, it can be argued that learning to problem solve is better developed by solving problems similar to those seen by experts rather than well-structured problems that have a single-solution.

Researchers have investigated approaches for developing novices' ability to think like an expert when solving problems. Examples of these approaches included giving learners the opportunity to develop their own knowledge through authentic and realistic tasks (e.g., Lesh & Kelly, 1994; Derry, Levin, & Schauble, 1995), mentoring or coaching (e.g., Barnett, 1995), using pedagogical techniques, such as scaffolding (e.g., Chen & Ge, 2006; Ge & Land, 2004; Henningsen & Stein, 1997), and using instructional activities, such as problem-based learning (e.g., Weiss, 2003) and projects (e.g., Binnie, 2002). Because many of the approaches are related to using problems seen by experts, the following section focuses on examining literature related to using ill-structured problems to develop students' thinking into more expert-like thinking.

2.3.1 Use of ill-structured problems to develop expert thinking. Ill-structured problems have been used to develop novices' thinking in a wide range of students, from high school students (e.g., Chin & Chia, 2005) to graduate students (e.g., Barrows, 2000). They have also been used in variety of disciplines, such as biology (e.g., Chin & Chia, 2005), medical school (e.g., Barrows, 2000), and statistics (e.g., Derry, Levin, Osana, Jones, & Peterson, 2000).

Researchers have argued that instructional support is critical for developing the skills that are needed for solving ill-structured problems. Types of instructional support found in the literature included

- scaffolding (Chen & Ge, 2006; Ge & Land, 2004; Jonassen, 1997);
- question prompts (e.g., procedural prompts, elaboration prompts, reflection prompts) (Ge & Land, 2004; Chen & Ge, 2006);

- expert modeling (e.g., observe how experts think, compare experts' thinking with their own thought processes while solving the problem, and then internalize the thought processes of the experts) (Chen & Ge, 2006; Barnett, 1995; Jonassen, 1997); and
- peer interactions (e.g., multiple perspectives and to have students see thing they might not have seen) (Ge & Land, 2004; Chen & Ge, 2006; Weiss, 2003).

Another instructional support method included scaffolding the trajectory of problem contexts. Woods et al. (1997) recommended that the skills needed to solve ill-structured problems should be developed first by using problems and activities that are context-independent. Then, these skills can be bridged to problems where the context is more domain-specific. The final step is to expand on the skills to problems in other contexts and in everyday life situations.

Aside from instructional support, designers of ill-structured problems also play a key role in successful development of the skills needed for solving ill-structured problems, according to Jonassen (1997). Table 2 describes Jonassen's six guidelines for designing an ill-structured problem. He also suggested that designers collaborate with subject matter experts and experienced practitioners while designing the problems. The information gathered from the experts and practitioners can help the designer in interpreting and understanding the problem-solving tasks (Jonassen & Hernandez-Serrano, 2002).

Table 2

Jonassen's (1997) Considerations and Reason for Consideration for Ill-Structured Problem Designers

Consideration	Reason for Consideration
1. Articulate problem context	The problem context needs to build on the problem solver's knowledge in order for prior problem-solving skills to be transferred.
2. Introduce problem constraints	The problem constraints need to be set up so that there is not a clear or obvious solution or solution alternative.
3. Locate, select, and develop cases for learners	The cases should be realistic situations with multiple solutions that are challenging, yet solvable in order for the necessary problem-solving skills to develop.
4. Support knowledge base construction	Knowledge construction occurs by having problem-solvers identify alternative opinions and perspectives and reconcile the multiple perspectives
5. Support argument construction	Argument construction aids in problem-solvers engaging in epistemic cognition and metacognitive thinking
6. Assess problem solutions	Assessment of the solutions should be based both the problem-solver's process and product.

2.3.2 Model-eliciting activities as ill-structured problems. One type of ill-structured problem that has been studied by mathematics and engineering educators is referred to as a model-eliciting activity (MEA). MEAs are described by Lesh et al. (2000) as

> thought revealing activities that focus on the development of constructs (models or conceptual systems that are embedded in a variety of representational systems) that provide the conceptual foundations for deeper and higher order understandings of many of the most powerful ideas in precollege mathematics and science curricula (p. 592).

While the initial purpose of MEAs was to help investigate and develop students' thinking during the activity, MEAs have been found to have broader instructional value as well (e.g., Diefes-Dux, Moore, Zawojewski, Imbrie, & Follman, 2004).

MEAs are designed to meet six principles to ensure that they reveal students' development of the construct. These six principles are the *model construction* principle, the *reality* principle, the *self-assessment* principle, the *model documentation* principle, the *model share-ability and reusability* principle, and the *effective prototype* principle (Lesh et al., 2000). It is also proposed that these six principles contribute to student's understanding of mathematical concepts, problem solving, metacognition, communication, and teamwork skills (e.g., Lesh & Doerr, 2003; Lesh & Zawojewski, 2007). Table 3 describes the six principles in detail.

Table 3

Lesh et al. (2000) MEA Design Principles

Design Principle	Description of Principle
1. Model construction principle	Students need to explicitly mathematize the problem by developing a model to interpret the givens, goals, and possible solution of the problem.
2. Reality or meaningfulness principle	MEAs are designed around contexts that require students to use their own personal knowledge and experiences in solving the problem.
3. Self-assessment principle	Due to the multiple modeling cycles students go through before coming up with a final product, students need to be able to assess how well their products meet the client's stated purpose.
4. Model documentation principle	MEAs are designed so that students are required to reveal explicitly their thinking processes. This is accomplished by having students document the givens, goals, and possible solutions that they considered while solving the problem and by having them work in groups.

5. Model share-ability and reusability principle	The model needs to be shareable, transportable, easily modifiable, and reusable to other similar situations in order for students to develop general ways of thinking.
6. Effective prototype principle	The solution needs to be a useful prototype or metaphor for interpreting other problems that are structurally similar; that is, the model created by the students will be as simple as possible, while still establishing the need for a better method to solve the problem.

MEAs are primarily used and studied in three fields of education: mathematics education, engineering education, and statistics education. In mathematics education, the research has mainly focused on middle school students (Lesh & Harel, 2003; Chan, 2008), whereas the research in both engineering and statistics education have mainly focused on undergraduate students (e.g., Bursic, Shuman, & Besterfield-Sacre, 2011; Moore & Hjalmarson, 2010; Carnes, Cardella, & Diefes-Dux, 2010; Lesh, Amit, & Schorr, 1997; Hjalmarson, Moore, & delMas, 2011). The results from the research on using MEAs in these three fields are now described.

2.3.2.1 Research on using MEAs in mathematics education. In mathematics education, researchers have identified positive outcomes as a result of using MEAs in the class. One positive outcome was the high cognitive and metacognitive demands that were placed on the students beyond the thinking that is required when solving traditional word problems (Lesh & Harel, 2003; Chan, 2008). This demand is most likely due to the ill-structured nature of the MEA problem. Consequently, by forcing students to confront their own superficial way of thinking about a concept during an MEA, Chan (2008) found that students develop a more sophisticated way of thinking about the concept.

Another positive outcome that has been documented is termed *local conceptual development* (Lesh & Kaput, 1988). This is the conceptual development process that students go through while solving an MEA for 60-90 minutes, which is comparable to the process that developmental psychologists have observed for the same construct over a period of several years. Lesh and Harel (2003) hypothesized that local conceptual development occurs because the students are (a) challenged to develop models and conceptual tools that are sharable, reuseable, and transportable; (b) introduced to powerful representation systems for expressing relevant constructs; and (c) encouraged to go beyond thinking with these constructs to also think about them. As a result, the researchers believed that the transferability of the constructs and conceptual systems are enhanced.

2.3.2.2 Research on using MEAs in engineering education. As previously mentioned, the research in engineering education has focused on undergraduate students in introductory engineering courses. When MEAs were used in the engineering classroom, positive results were found in the following areas: (a) the development of professional skills as described by the Accreditation Board of Engineering and Technology (ABET) (Bursic et al., 2011), (b) the ability to solve complex engineering problems (Moore & Hjalmarson, 2010), and (c) the ability to change students' thinking by incorporating peer and TA feedback into the activity (Carnes et al., 2010).

2.3.2.3 Research on using MEAs in statistics education. Similar to the research in mathematics education and engineering eduation, the MEAs studies in statistics education have also reported positive outcomes. Lesh et al. (1997) found that students'

thinking about statistical concepts evolved from an informal, uncoordinated way of thinking (e.g., simple calculations without interpreting or mathematizing the data) toward more formal ways of thinking (e.g., trends, measures of tendency, graphical analysis and representations of the data) while solving a statistical MEA. Another study found that MEAs had instructional benefit. Because students struggled with implementing key statistical concepts in an MEA about sampling and variability, Hjalmarson et al. (2011) suggested that instructors could use the MEA to help identify students' misconceptions about measures of center and variability.

2.3.3 Summary of strategies to develop expert thinking. Experts commonly encounter ill-structured problems within their field. As a result, presenting students with problems similar to those seen by experts (i.e., ill-structured problems) was one suggestion for developing expert-like thinking in students. A type of ill-structured problem that has been used in mathematics, engineering, and statistics education is MEAs. Both ill-structured problems and MEAs have recommended design principles that should be followed to ensure the development of expert-like thinking in students. Positive results in students' thinking and problem-solving skills have been found when ill-structured problems and MEAs were used to develop their expert-like thinking (e.g., domain-general problem-solving skills, local conceptual development, metacognitive skills, ABET professional skills, a more formal way of thinking).

To understand whether the strategies for developing expert thinking are effective, methods for assessing thinking need to be considered. Literature related to assessing

thinking from the general area of ill-structured problems and the specific areas of MEAs and statistical thinking are examined next.

2.4 Assessing Expert Thinking

To assess how problem-solvers think through a problem, cognitive and metacognitive processes need to be extracted. The following section examines how thought-processes during problem-solving situations have been assessed in the areas of ill-structured problems, MEAs, and statistical thinking.

2.4.1 Use of ill-structured problems to assess thinking. Researchers have collected a variety of data formats to investigate the thinking that occurs during ill-structured problems. These included written responses from assessments with ill-structured problems (e.g., Schraw et al., 1995; Heller, Keith, & Anderson, 1992; et al., 2003), verbal descriptions from think-aloud protocols or interviews while solving ill-structured problems (e.g., Chen & Ge, 2006; Ho, 2000; Fernandes & Simon, 1999; Voss, 2006; Derry et al., 2000; Ertmer et al., 2008; Chin & Chia, 2005), and graphical representations from an ill-structured problem scenario (Spector, 2006).

Similar to the data formats that were collected, multiple methods for assessing problem-solvers' thinking while solving ill-structured problems were found in the literature. One method was using an existing framework on reasoning and thinking to code the qualitative data (e.g., Schraw et al., 1995; Voss, 2006; Ho, 2000, Fernandes & Simon, 1999). Other examples in the literature were using of an exploratory qualitative analysis approach to find common themes or patterns in the qualitative data (e.g., Derry et al., 2000; Ertmer et al., 2008; Chin & Chia, 2005) and devising a scoring scheme for

the assessments (Heller et al., 1992; Shin et al., 2003; Derry, et al., 2000). These scoring schemes were based on expert-like problem-solving characteristics and on responses similar to expert-like responses.

Due to the nature of the data formats and methods in the studies in this area, multiple coders were commonly used to examine the reliability of their coding, that is, to show consistency among multiple coders (Derry et al., 2000; Schraw et al., 1995; Shin et al., 2003; Ertmer et al., 2008; Fernandes & Simon, 1999; Chin & Chia, 2005). However, one study was found that used only one coder (Ho, 2000). To improve the consistency and objectivity of the encoding process in this study, Ho (2000) coded the data twice, with a month separation between encoding tasks to minimize the influence of the first coding process.

2.4.2 Use of MEAs to assess thinking. Similarities and differences were found between the studies that used ill-structured problems and the studies that used MEAs with respect to methodologies for analyzing the data. For example, both areas of study had similar data formats: verbal transcripts from audiotape and videotapes, written responses, or a combination of the two (e.g., Lesh & Harel, 2003; Moore, Miller, Lesh, Stohlmann, & Kim, 2013; Carlson, Larsen, & Lesh, 2003). However, unlike the research on ill-structured problems that need think-aloud protocols to gather verbal descriptions of the problem-solver's thinking process, the thought-revealing process is a natural part of the MEA (Diefes-Dux et al., 2004). English, Lesh, and Fennewald (2008) argued that the documentation trail provided by the written solutions to the MEA could supplement the information obtained from video analyses.

Studies that used MEAs to assess thinking also differed in their unit of analysis, as compared to research on ill-structured problems. Lesh and Zawojewski (2007) posited that researchers could use students who solve problems in groups to help understand students who solve problems individually. While working in groups to solve a problem, it is natural for the internal thinking processes and ideas to become externalized through various outlets (e.g., talking, writing, drawing pictures, expressing, testing, revising solutions). Consequently, they argued that even though the unit of analysis is different (i.e., groups verses individuals) between the two types of studies (i.e., expert-novice verses MEAs), the goals between studies were similar; that is, both types of studies focus on how mathematical ideas develop and on the attributes that prompt rethinking and revision of those ideas.

Lastly, studies that used ill-structured problems and studies that used MEAs had similar approaches to evaluating and scoring the data. The characteristics of students' models were evaluated using rubrics (see Table 4) (e.g., Diefes-Dux et al., 2004, Diefes-Dux, Zawojewski, & Hjalmarson, 2010; Hjalmarson et al., 2010), a preexisting framework (e.g., Moore et al., 2013; Carlson et al., 2003), or a qualitative analysis approach with the goal of finding common themes or patterns in the models (e.g., Hjalmarson et al., 2011). These characteristics were also often evaluated using triangulation of multiple data sources and multiple coders to provide trustworthiness of the results and transparency of the data analysis (e.g., Moore et al., 2013).

The type of evaluation technique used in these MEA studies (e.g., rubric, pre-existing framework, or qualitative method) depended on the purpose of study. If the

purpose for using MEAs was to understand whether students were building the conceptual foundations that are essential for a topic, then researchers tended to use rubrics (see Table 4) (e.g., Diefes-Dux et al., 2004, Diefes-Dux, et al., 2010; Hjalmarson et al., 2010). These rubrics were beneficial to both researchers and students.

Table 4

Summary of Rubrics Used on MEA Solutions to Assess Thinking

Article	Purpose of Rubric	Scoring Scheme
Diefes-Dux et al. (2004)	To assess the quality of the students' work as well as the degree to which they met the client's needs and could be generalized to similar situations.	Scale: • 1 = Clear, concise, and useful • 0.5 = Requires minor edits • 0 = Is non-existent or requires major editing Two scores: • Individual contribution • Team contribution
Diefes-Dux et al. (2010)	To assess the characteristics of appropriateness of the mathematical model, attention to audience, and generalizability of the product that were deemed valuable by engineering experts.	Scale: • 0, 1, 2, 3, 4, with 4 = fully demonstrates conceptual understanding in the specific MEA, and 0 = does not demonstrate conceptual understanding in the specific MEA Three scores: • Appropriateness of the mathematical model • Attention to audience • Generalizability of the product
Moore & Hjalmarson (2010) & Hjalmarson et al. (2011): Quality Assurance Guide	To assess how well the procedure can be generalized or used in another situation, whether the client's needs were met, and whether there are clear directions on how to use the	Scale: • 1 = Requires redirection • 2 = Requires major extensions or revisions • 3 = Requires editing and revisions

(QAG)	procedure.	• 4 = Useful for this specific data given, but not shareable or reusable OR Almost shareable and reusable but requires minor revisions • 5 = Shareable and reusable One score: • Usefulness of product
Yildirim, Shuman, and Bestefield-Sacre (2010)	To assess how well the solution achieves or executes the following MEA principles: generalizability, self-assessment/testing, model documentation, and effective prototype.	Scale: • 1 = Principle was not achieved or executed • 2 = Principle was somewhat achieved or executed • 3 = Principle was sufficiently achieved or executed • 4 = Principle was achieved or executed in a good manner • 5 = Principle was achieved or executed in an outstanding manner Average of four scores: • Generalizability • Self-assessment/testing • Model documentation • Effective prototype

In contrast, if the purpose for using MEAs was to understand how students develop the conceptual foundations for a topic, then researchers tended to use preexisting frameworks or qualitative methods (e.g., Carlson et al., 2003; Moore et al., 2013; Moore & Hjalmarson, 2010; Hjalmarson et al., 2011; Chamberlin, 2004; Chamberlin, 2005). These more exploratory evaluation techniques helped to revise existing theories on how students' solve problems (Carlson et al., 2003; Moore et al., 2013), understand students' stages for developing a model and students' transitions through those stages (Moore &

Hjalmarson, 2010; Hjalmarson et al., 2011), and assist teachers in developing a sheet that explained students' thought processes when solving an MEA (Chamberlin, 2004; Chamberlin, 2005). However, regardless of the evaluation technique, Lesh & Zawojewski (2007) argued that

> one of the most significant characteristics of models-and-modeling perspectives on mathematics learning and problem solving is the assumption that the theory, model, or research perspective being used by the researcher needs to be expressed, tested, and revised systematically as part of the research process (p. 797).

Therefore, as suggested by Lesh and Zawojewski, the researcher's model of students' thinking should evolve within and between studies.

Researchers were not the only ones that benefited from the thinking that occurs during an MEA. Students also benefited from assessing and reflecting on their thought processes while solving an MEA, which forces students to use metacognitive processes. For example, in a follow-up homework, students compared and contrasted their final solution with a professional method of analysis (Diefes-Dux et al., 2004). This assessment forced the students to evaluate their model against other competing models. Another example of explicitly having students reevaluate their thinking was using a model-adaptation activity following an MEA (Lesh, Cramer, Doerr, Post, and Zawojewski, 2003). This activity was suggested a way to give students the opportunity to adapt their own model or another model to a new situation.

To summarize the literature on assessing thinking in ill-structured problems and MEAs, a variety of methods have been used to assess students' thinking. Similarities between the two areas included similar data formats, evaluation techniques, multiple coders, whereas the two areas differed in their unit of analysis (i.e., group vs. individual). To further build on understanding how to assess thinking, the next section presents literature related to assessing an expert way of thinking specific to statistics: statistical thinking, as defined by Wild and Pfannkuch's framework of statistical thinking.

2.4.3 Assessing statistical thinking. As previously mentioned, statistical thinking is often used interchangeably with the terms statistical literacy and statistical reasoning. The exchangeability of terms could be attributed to having no accepted definition of the three terms. However, delMas (2002) argued that it is the nature of the assessment that helps distinguish whether an item is measuring statistical literacy, reasoning, or thinking. He stated

> statistical thinking is promoted when instruction challenges students to apply their understanding to real world problems, to critique and evaluate the design and conclusions of studies, or to generalize knowledge obtained from classroom examples to new and somewhat novel situations (An Alternative Perspective section, para. 2).

Aligned with delMas's idea of assessing statistical thinking, researchers have developed end-of-the-course assessment items to try to elicit students' statistical thinking (Chance, 2002; Garfield, et al., 2012). These free-response items were constructed to elicit thinking about the whole statistical investigative process, rather than just focusing

on a specific statistical procedure. Chance (2002) proposed that in order "to determine whether students are applying statistical thinking, problems need to be designed that test student reflexes, thought patterns, and creativity in novel situations" (Conclusion section, para. 3). She used Wild and Pfannkuch's framework of statistical thinking as the basis for assessing students on the needed mental habits and problem-solving skills for thinking statistically.

2.4.3.1 Using Wild and Pfannkuch's framework to assess statistical thinking.

Aspects of Wild and Pfannkuch's statistical thinking framework have been used to assess students' or teachers' thinking when solving statistical problems. Groth (2005) used the relationship between the contextual knowledge and statistical knowledge from Wild and Pfannkuch's framework to help interpret the patterns of students' thinking within two statistical contexts, signal-verses-noise and typical value. Fifteen students were interviewed and asked to solve problems that involved thinking about typical values and about signal-versus-noise situations. The student responses from the interview data were analyzed qualitatively by grouping strategies for solving each context with other strategies that shared common characteristics. He found evidence that students who exhibited a "high level of statistical thinking" (p. 122) when solving problems about typical values and signal-verses-noise contexts continually moved back and forth between the data and the context of the problem.

Makar and Confrey (2002) researched teachers' statistical thinking when solving a statistical problem of comparing two groups. To assess teachers' levels of statistical thinking, they developed a hierarchical taxonomy based on elements from Wild and

Pfannkuch's statistical thinking framework. The elements used in their taxonomy from Wild and Pfannkuch's framework included "anticipation of variation, ability to construct and use models, good statistical and contextual knowledge base, and ability to synthesize these elements to produce conjectures and inferences" (p. 3). Using teachers' responses from the pre-post assessments on a comparing two-groups situation, the researchers categorized the teachers into the taxonomy levels. Makar and Confrey concluded that of the four teachers that remained in the study, two of the teachers regressed in the levels of the taxonomy. Additionally, they found that teachers who left the study started with weaker content knowledge than those who remained. Based on their results, Makar and Confrey advocated for creating real-life statistical experiences for teachers to modify teachers' understanding of data analysis.

Another study investigated the beginning stages of statistical thinking for 12 secondary students' while they explored a small multivariate dataset (Pfannkuch & Rubick, 2002). The researchers qualitatively analyzed the students' thinking during their exploration by using the statistical thinking framework proposed by Wild and Pfannkuch. In particular, the researchers focused on the types of thinking that were fundamental to statistics: transnumeration, consideration of variation, reasoning with statistical models, and integration of the contextual and statistical. From this exploratory study, five aspects were identified that can be used for determining how students think, interpret, and understand multivariate data. These include using prior contextual and statistical knowledge, thinking at a higher level than constructed representations, actively representing and construing the data, shuttling back and forth between local thinking and

global thinking, and using statistical thinking differently across the various data representations.

Other researchers have used Wild and Pfannkuch's framework of statistical thinking to look for evidence of statistical thinking within secondary classroom settings (e.g., Pfannkuch & Horring, 2005; Pfannkuch & Rubick, 2002). Pfannkuch and Horring (2005) investigated secondary classrooms that used a curriculum based on Wild and Pfannkuch's framework as its underlying theoretical model for developing students' statistical thinking. To find evidence of statistical thinking within the classrooms, the data collected were qualitatively analyzed by classifying the data into Wild and Pfannkuch's framework. Pfannkuch and Horring found that students' statistical thinking were being developed throughout the implemented curriculum, such as beginning to use the investigative cycle when solving statistical problems. However, the researchers also noticed that elements of the statistical thinking model were not being incorporated into the implemented curriculum (e.g., drawing conclusions). Pfannkuch and Horring predicted that with more time and experience, teachers would learn to integrate more statistical thinking elements into the curriculum.

2.4.4 Summary of the assessment of expert thinking. In summary, similar types of data were collected to assess expert thinking across the areas of ill-structured problems, MEAs, and statistical thinking. These types of assessment included verbal transcripts from thinking-protocols, interviews, videotapes, or audiotapes, and written solutions to assessment tasks. Researchers using MEAs to assess expert thinking argued that the information gleaned from students' written solutions and videotape transcripts

provide similar information as expert-novice studies because of the design of the model-eliciting activity. Suggestions for how to elicit statistical thinking from the statistical thinking literature included creating free-response items that get students to think about the overall statistical investigative process and to challenge students' mental habits. It was also suggested that these items should integrate the rote facts, formulas, and statistical knowledge into the context of a real-world problem.

The studies reviewed on expert thinking relied heavily on qualitative data and analyzed it using existing frameworks on thinking or exploratory methods to find common themes of students' cognitive processes while solving a problem. Additionally, the use of multiple sources of data and multiple coders was commonly used. In the MEA literature, when rubrics were used to score students' solutions, they were both to inform researchers and provide feedback to students on their learning. It was suggested in the MEA literature that researchers should test and revise their model of thinking by using students' responses as part of the research process.

Many researchers assessing statistical thinking have used Wild and Pfannkuch's framework in multiple ways. One researcher looked at a single aspect of their framework (Groth, 2005) whereas others incorporated multiple aspects of their framework into their assessment (Chance, 2002; Makar & Confrey, 2002; Pfannkuch & Horring, 2005; Pfannkuch & Rubick, 2002). Wild and Pfannkuch's framework was also used to examine the thinking that occurs during a single statistical problem or in the wider arena of a curriculum that had a learning objective of develop students' statistical thinking.

One key difference between the three areas of literature on assessing expert thinking (i.e., ill-structured problems, MEAs, and statistical thinking) was the unit of analysis under investigation. In the ill-structured problems literature and statistical thinking literature, the unit of analysis was an individual problem-solver, whereas the unit of analysis in the MEA literature was often a group of problem-solvers.

2.5 Discussion

Many statistics courses have a learning objective of develop students' statistical thinking. If this is an important learning outcome, then it must be operationally defined so that statistics education researchers and instructors can work towards developing and assessing it. A critical review of the literature was presented to understand how statistical thinking has been defined, how using ill-structured problems can be used to develop expert thinking (e.g., statistical thinking) in novices, and how this thinking can be assessed. In this section, a summary and critique of the literature is presented and implications for developing and assessing statistical thinking are given.

2.5.1 Summary and critique of the literature.

From the perspectives in the areas of total quality improvement, statistics, and statistics education, statistical thinking is generally described as thinking used by expert applied statisticians when solving statistical problems. However, only Wild and Pfannkuch, from the perspective of statistics education, conducted empirical research to construct a framework on the dimensions of thinking like a statistician. Their research resulted in developing a framework about statistical thinking, as opposed to listing a few elements of statistical thinking as seen in perspectives from Snee and Moore. Based on Wild and Pfannkuch's

research, it appears that statistical thinking is more a complex process than a list of four or five elements.

One critique of Wild and Pfannkuch's framework of statistical thinking was that it did not include research from other fields that investigate thinking of experts. In light of this, literature on understanding what it means to think like an expert in mathematics and in more domain-general fields were reviewed to expand on Wild and Pfannkuch's framework. This review found similar descriptions of what it means to be an expert between the two areas (i.e., mathematical thinking and expert-novice literature). For example, both expert mathematicians and experts in the expert-novice literature were found to easily transfer knowledge to new problems, identify important cues within a problem, use metacognition, apply knowledge in a flexible manner, and have conditionalized knowledge and deep knowledge structures.

To provide a more complete understanding of what it means to think statistically, the characteristics of expert thinking from the areas of mathematical thinking and expert-novice research are now compared to Wild and Pfannkuch's framework of statistical thinking (see Table 5). The majority of Wild and Pfannkuch's framework were similar to the characteristics found in the expert literature. For example, the investigative cycle (Dimension 1), two-thirds of the elements from the types of thinking cycle (Dimension 2), and the interrogative cycle (Dimension 3) from Wild and Pfannkuch's framework could be mapped to general expert thinking characteristics. Differences were found, as well, between the framework and the expert literature. Based on this comparison, a more

complete conceptualization of the components of statistical thinking should integrate both sets of characteristics.

Table 5

Comparing and Contrasting Expert Thinking Characteristics to Wild and Pfannkuch's (1999) Framework

	Expert Thinking Characteristics	Wild and Pfannkuch's (1999) Framework
Similarities		
	• Adaptive expertise	• Transnumeration (Element of Dimension 2)
	• Organization of knowledge (i.e., ability understand and represent problems; conditionalized knowledge)	• Investigative cycle (Dimension 1) • Consideration of variation (Element of Dimension 2) • Reasoning with statistical models (Element of Dimension 2) • Strategic (Element of Dimension 2) • Applying techniques (Element of Dimension 2)
	• Metacognitive skills	• Interrogative cycle (Dimension 3) • Modeling (Element of Dimension 2)
Differences		
	• Fluent and automatic retrieval of knowledge • Problem-solving techniques o Skipping steps; collapse multiple steps o Working forward approach to solving problems	• Recognition of the need for data (Element of Dimension 2) • Integrating the statistical and contextual (Element of Dimension 2) • Seeking explanations (Element of Dimension 2) • Individual dispositions (Dimension 4)

The review of the literature on developing expert thinking suggested that learning to think like an expert is not a natural learning process. It takes considerable amount of time, energy, and particular dispositions to develop expert thinking within a domain. Therefore, if developing statistical thinking is a learning objective for a statistics course, then an expertise-building approach to learning needs to be incorporated into the course. To this end, literature on ill-structured problems was reviewed to provide guidelines for how to develop expert thinking in students by using problems frequently encountered by expert statisticians (i.e., ill-structured problems). Guidelines from the literature included

- using realistic situations that are encountered within the domain,
- having multiple solutions without a prescribed solution path,
- building on problem-solvers existing knowledge to transfer prior problem-solving skills to the context,
- requiring problem-solvers to evaluate their solutions through an iterative process of monitoring and refining the solutions until they feel their solution is "correct",
- requiring problem-solvers to provide justification for their solutions,
- having instructional support during the ill-structured problem, and
- using real-world criteria for assessing the effectiveness of the solutions and problem-solving processes (e.g., constructing an argument by using evidence to support their solutions, reusability of solutions to a similar situation).

Similar guidelines for how to develop expert thinking were seen in the literature on MEAs. Recall that an MEA is an ill-structured problem unique to the disciplines of

mathematics education, engineering education, and statistics education. The MEA research suggested that MEAs help develop students' thinking. Due to the natural link between mathematics and statistics, MEAs could be a useful activity to develop students' statistical thinking in statistics courses. During an MEA, students (i.e., novices) are forced to revise and reassess their thinking of the concept. This metacognitive process aids in building and developing their thinking into a more expert-type of thinking, such as statistical thinking.

MEAs could also be useful because the activity uses a problem from a real situation. Research has found that students who are highly interested in a problem are more motivated to learn the appropriate skills needed to solve similar problems (e.g., Schiefele, 1999). In this regard, MEAs use real-world contexts that are often of interest to students. Using a context that students are interested in can help even the most frightened student become invested in solving a statistical problem while simultaneously develop their thinking during the activity (e.g., Colvin & Vos, 1997; Bereiter & Scardamalia, 1986). However, these arguments of using MEAs to develop students' statistical thinking are speculative. There is no empirical evidence about whether MEAs develop students' statistical thinking. Additionally, there is no research on how the thinking that develops during the MEA transfers to similar types of problems.

The review of the literature on assessing expert thinking revealed that assessing the thought processes of a problem-solver appears to be a challenging and time-consuming process. The data collected were qualitative via verbal transcripts and written solutions to a problem. Qualitative data, as opposed to quantitative data, are typically

more difficult to analyze. This data was then analyzed using a rubric, a preexisting framework on thinking, or exploratory methods to find common patterns of thinking. They also used triangulation of multiple sources of data and multiple coders as recommended by qualitative researchers (Guba, 1981).

For the studies that assessed students' statistical thinking, researchers frequently used Wild and Pfannkuch's framework of statistical thinking as the basis for their assessment. However, the majority of these studies did not incorporate all elements of the framework into the assessment. The researchers of these studies, instead, selected particular aspects of the statistical thinking framework to help evaluate a portion of the statistical thinking construct. The one study that used all aspects of Wild and Pfannkuch's framework assessed students' thinking in the context of a curriculum, rather than in a small-scale problem. Consequently, more studies are needed to assess the complete construct of statistical thinking in their students.

Based on suggestions for how to assess expert thinking and statistical thinking, MEAs could be used to also assess students' statistical thinking. One benefit of using MEAs in the classroom is it is an assessment as learning, that is, "students engage in new learning by monitoring and adapting their own understanding via the assessment process" (Garfield & Franklin, 2010, p. 134). MEAs can be completed during a 50-60 minute class period and have a final product that contains an audit trail of the groups' thinking processes. However, there was no consensus in the MEA literature on what or how to assess in the students' final product of an MEA. Therefore, it is unclear how instructors should assess the students' final product or how the feedback from the instructor can

benefit students' development of their thinking. To this end, quality assessments, potentially using MEAs as their basis, are needed to measure students' statistical thinking.

2.5.2 Critique of the literature on developing and assessing statistical thinking.

As previously mentioned, many statistics courses have a learning objective of develop students' statistical thinking. However, based on this literature review, there appears to be little evidence about how to develop students' statistical thinking and assess this development.

The literature suggested the use of instructional supports, such as scaffolding, peer review, mentoring, and question prompts, to develop expert-like thinking in novices. Research is needed to investigate the use of supports for developing students' statistical thinking. Examples of instructional supports that could be investigated include the use of instructor modeling of thinking, through discussion, scaffolding, or question prompts in activities, or the use of cooperative learning environment in a statistics course.

The literature included suggestions regarding the use of ill-structured problems, such as MEAs, for developing and assessing students' thinking. In the previous section, the review suggested that MEAs could be used to develop and assess students' statistical thinking in a statistics course. Additional research is needed to explore the role of MEAs in developing and assessing students' statistical thinking in a statistics course. Example areas of research that can be investigated include the number of MEAs in the course (e.g., one at the beginning of each unit or many throughout each unit) and the use of MEAs in

the course (e.g., as an assessment of statistical thinking or as an activity to prime the field for a concept).

The word *developing* implies causing something to grow or become more advanced over time. Therefore, a longitudinal study is needed to assess the development of students' statistical thinking in a statistics course or within a lesson. To this end, quality assessments are needed that evaluate students' statistical thinking at multiple times in a course. These assessments would help instructors understand how their students are developing statistical thinking and could suggest changes in curriculum and pedagogy.

2.5.3 Problem statement. Currently, statistics courses with the learning objective of "develop students' statistical thinking" are not able to assess whether they have met their goal because there is no known assessment that measures the complete construct of statistical thinking. To fill this gap, new quality assessments that measure statistical thinking need to be created. These assessments should use ill-structured problems (e.g., MEAs) to assess the expert-like thinking that occurs while solving problems. Additionally, to measure the development of students' statistical thinking, these assessments need to be given multiple times in a statistics course to understand whether students' statistical thinking is being developed. Therefore, the aim of this study, as outlined in the following chapter, is to develop an assessment of statistical thinking to understand what components of students' statistical thinking is revealed and developed in an introductory course that is based on modeling and simulation.

Chapter 3

Methods

3.1 Introduction

This study aimed to answer the following research question:

What components of students' statistical thinking are revealed and developed in an introductory statistics course that is based on modeling and simulation?

To answer this research question, an assessment was developed and then administered at the beginning and end of an introductory statistics course. This introductory statistics course has an explicit learning objective of "develop students' statistical thinking." This chapter describes the development, administration, and analysis of this assessment of statistical thinking, called Modeling to Elicit Statistical Thinking (MODEST).

3.2 Overview of the Study

This study is composed of three phases. The first phase involved developing MODEST. This included writing items around a MEA context using the test blueprint, collecting reviewer feedback on MODEST, and making modifications to MODEST based on the reviewer feedback.

The second phase of this study was pilot testing MODEST to two cohorts of students. The first cohort consisted of senior, undergraduate students majoring in statistics. This cohort was chosen because they should be able to think statistically. These senior statistics students participated in cognitive interviews and their responses were used to determine whether or not the items in MODEST elicited statistical thinking. Based on the data from these interviews, another version of MODEST was created. The

second cohort was students enrolled in an introductory statistics course during the Fall 2014 semester. This cohort was chosen because they were representative of the student population that was used for the field test phase of the study. The students in this introductory statistics course were administered MODEST as an online assessment at the end of the semester. Similar to the senior statistics students, the responses from these students were examined to determine whether or not the items in MODEST elicited statistical thinking. Based on the data from this online administration, the final version of MODEST was created.

The third phase of the study, known as the field test, consisted of administrating MODEST to students enrolled in an introductory statistics course during the Spring 2015 semester. These students completed the online assessment twice, once at the beginning of the semester (Pre administration) and once at the end of the semester (Post administration). The student responses were compared between the Pre administration and the Post administration to answer this study's research question. The three phases–assessment development, pilot test, and field test–are described in detail in the next sections.

3.3 Assessment Development

The development of the assessment followed a construct-centered approach of assessment design, as proposed by Messick (1994).

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or

> are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics" (p. 16).

Following this approach, a test blueprint was first created to identify the characteristics of statistical thinking that would be measured by MODEST. Then, items were written to elicit the statistical thinking identified in the test blueprint. The test blueprint was also used to analyze the data collected in the study. This assessment development process is now described.

3.3.1 Test blueprint. A test blueprint was created by integrating the expert characteristics from the expert-novice literature (e.g., Bransford, et al., 2000) with Wild and Pfannkuch's framework of statistical thinking (1999). The domain-specific and domain-general expert characteristics were merged to provide a more complete understanding of what it means to think like an expert within the domain of statistics. As a result, the test blueprint consists of four components of statistical thinking:

- General Problem-Solving Characteristics,
- Statistical Problem-Solving Processes,
- Cognitive Processes of Statistical Problem-Solving, and
- Individual Dispositions.

These components of statistical thinking are made up of elements of statistical thinking (see Table 6). For example, the component of Statistical Problem-Solving Processes is

made up of three elements: *develops a plan for collection or analysis of the data*, *analyzes the data*, and *draws a conclusion*.

As previously mentioned, this test blueprint was then used to write items and analyze the data collected in the study. The data collected included feedback from reviewers and students' responses to MODEST. After each data collection, the test blueprint was revisited and updated to ensure that it was capturing the statistical thinking that was elicited by the items in MODEST (see Appendices B1 to B4 for all of the test blueprints). These revisions are described in detail in the Results chapter. The final test blueprint for MODEST is presented in Table 6.

Table 6

Final Test Blueprint for MODEST

Components Of Statistical Thinking	Item Number
General Problem-Solving Characteristics	
• Creates a model o Produces a conceptual model o Translates the conceptual model into a statistical model o Quality of the model *Description: Creates a model to help understand and predict a real-life situation.*	• Item 4 o Produces a conceptual model • Item 5: o Translates the conceptual model into a statistical model o Quality of the model
Statistical Problem-Solving Processes	
• Develops a [reasonable] plan for collection or analysis of the data, which includes: o Identifying types of information that is needed o Seeking information using pre-existing knowledge or an external source.	• Item 1 • Item 3 • Item 13
• Analyzes the data *Description: Fits and assesses a model to solve the*	• Item 7

problem. • Draws a conclusion *Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.*	• Item 10: o Interprets findings.
Cognitive Processes of Statistical Problem-Solving	
• Considers variation *Description: Includes:* o *Explaining variation among variables or cases.* o *Looking for sources of variability by examining patterns in the variables or relationships between variables.* o *Considering measurement error.*	• Item 2 • Item 8
• Reasons with statistical models *Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics.*	• Item 9 • Item 10 • Item 13
• Recognizes the need for data	• Item 13
Individual Dispositions	
• Is curious *Description: Is curious by asking Items such as, "is this something that happens more generally?"*	• Item 12
• Is skeptical (e.g., is this conclusion justified?)	• Item 10 • Item 11

This test blueprint does not include all of the characteristics of statistical thinking that were identified in the literature. First, characteristics related to complex cognitive processes (i.e., automatic and fluent retrieval of knowledge, organization of knowledge, and metacognition) were not assessed because these processes are typically assessed via verbal reports, observational techniques, and writing activities (e.g., Pearson, 2011), which were beyond the scope of this study. Characteristics related to problem

representation were not assessed because the problem context used in MODEST was ill-structured, but not ill-defined. That is, students were familiar with the context of the problem and did not need to try to understand the task for solving the problem. The individual disposition characteristics of perseverance and engagement were also not assessed because they were not of interest for this study.

3.3.2 Item development. As recommended in the literature on assessing "expert-like" thinking, a novel ill-structured problem was used as the foundation for MODEST. This novel problem was an MEA with a statistical focus, the Study Effectiveness MEA ("Measuring Study Effectiveness", 2013) (see Appendix A1). The primary task in the Study Effectiveness MEA is to create a summary score of study effectiveness for a survey about study effectiveness. The final product of this MEA is a written report that describes

(a) a method for assigning scores to respondents of the study effectiveness survey,

(b) an example of how to use the method on fake student responses to the survey, and

(c) a rank ordering of the fake students' scores of study effectiveness.

This MEA was chosen for several reasons. The first reason was no prior knowledge was required to solve the problem. Additionally, the problem scenario was relevant and familiar to students. Another reason was this MEA was one of a few statistical MEAs that was not seen in the introductory statistics course used in the field test. Most importantly, this MEA was chosen because it would elicit students' statistical thinking (e.g., *creates a model*, *considers variation*).

To create an assessment that measures individual students' statistical thinking, the MEA needed to be adapted from a group activity to an individual assessment. To do this, items were written around the Study Effectiveness MEA context and used the test blueprint as the basis for the statistical thinking elements that were assessed. Items were also developed with the intention of mirroring how an expert statistician would work through a problem similar to the problem in the Study Effectiveness MEA. The item format chosen for these items was constructed-response, which is the recommended item format for assessments that are trying to measure cognitive processes (Haladyna & Rodriguez, 2013). Other item writing guidelines from the literature were also considered, including carefully wording items so that the desired response was produced (Hogan & Murphy, 2007) and using positive wording in the items (Haladyna, Downing, & Rodriguez, 2002). The result of this item development process was the first version of MODEST (MODEST 1) and a test blueprint for MODEST 1 (see Appendices A2 and B1, respectively).

3.3.3 Assessment review and revision. Three sets of feedback on MODEST were gathered in this study. After each set of feedback, changes were made to MODEST to improve it as an assessment of statistical thinking. Additionally, all of the changes made were discussed with my advisors. The following sections describe the details of the reviewer feedback process.

3.3.3.1 Feedback from Statistics Education graduate students. The first set of feedback was gathered from three graduate students in the Statistics Education department at the University of Minnesota. They were instructed to complete MODEST 1

from an introductory statistics student's point of view. The primary purpose of this task was to gather preliminary evidence that students' responses to the items were revealing the statistical thinking identified in the test blueprint. The secondary purpose was to gather feedback related to the administration of the assessment. Based on the data collected from these graduate students, a second version of MODEST was created (MODEST 2).

3.3.3.2 Feedback from external reviewer. The second set of feedback was gathered from an external reviewer. This reviewer has a Ph.D. in Literature and was asked edit MODEST 2, particularly on the grammar and the clarity of the items. Based on her feedback, MODEST was modified for a third time (MODEST 3).

3.3.3.3 Feedback from expert reviewers. The final set of feedback was collected from five well-known and highly regarded statistics educators from different institutions and different countries. The reviewers' backgrounds consisted of statistical thinking (Reviewers 3 and 5), assessing students' statistical thinking (Reviewer 1), assessments (Reviewer 2), and a mathematics educator (Reviewer 4).

These expert reviewers were recruited via an email invitation (see Appendix C1), which was loosely based off of a reviewer invitation letter from another study (Park, 2012). They were asked to evaluate MODEST as an assessment of statistical thinking, as described by the test blueprint. The reviewers were provided with the following materials: MODEST 3 (see Appendix C2), the test blueprint (see Appendix C3), and an evaluation form (see Appendix C4).

Feedback from the expert reviewers was gathered at the item-level and at the assessment-level via the evaluation form. To obtain feedback at the item-level, data were collected via survey responses to the following question: "How much do you agree or disagree with the following statement? The assessment item measures the specified statistical thinking element." The reviewers marked "Agree", "Agree, but with reservations", or "Disagree" for all of the elements of statistical thinking within an item. They were also given the option to provide comments that explained their agreement rating. Then, data at the assessment-level were collected via survey responses to the following question: "How much do you agree or disagree with the following statement? Overall, the assessment appears to measure statistical thinking, based off of the test blueprint." The reviewers again marked "Agree", "Agree but with reservations", or "Disagree" and provided comments to explain their agreement choice. Both the item-level and assessment-level feedback were used to provide evidence of MODEST as an assessment of statistical thinking and to help improve MODEST and the test blueprint (see Appendix A3 for MODEST 4 and see Appendix B2 for its test blueprint).

3.4 Pilot Test

During the pilot test phase of the study, MODEST was administered to two cohorts of students. The first group was senior, undergraduate students majoring in statistics. This population of students was chosen because they should exhibit a more advanced way of thinking statistically than novices but not have the complete expert-like thinking of a statistician. The second group was students enrolled in an introductory statistics course, called the CATALST course, during the Fall 2014 semester. This course

has an explicit learning objective of "develop students' statistical thinking" and uses modeling and simulation-based methods to try to meet this objective (see Section 3.5.1 for more detail on the CATALST student population). The pilot CATALST students were chosen because they are representative of the student population in the field test. These two cohorts and their administrations of MODEST are now described.

3.4.1 Pilot test: Senior statistics students. The senior students, who were majoring in statistics, were recruited from an upper-level undergraduate statistics course (STAT 4893W: Senior Project) during the middle of Fall 2014 semester. Appendix F1 contains the script that used for recruitment. Interested students were contacted and in the end, four senior statistics students agreed to be a part of this study.

Data from these senior statistics students were collected via cognitive interviews. In the interview, the students talked through their thought processes while responding to the items on MODEST 4. This think-aloud approach is the recommended method when trying to understand the cognitive processes that occur when participants solve problems (van Someren, Barnard, & Sandberg, 1994). The materials used during the interview included an interview script (see Appendix F2), a consent form (see Appendix F3), and MODEST 4 (see Appendix A3). Based on the students' responses during the cognitive interview, the fifth version of MODEST was created.

3.4.2 Pilot test: CATALST students. The CATALST students were recruited during the final weeks of the Fall 2014 semester. Appendix F4 contains the script that used for recruitment. Following the in-class presentation, the instructors of the CATALST course were sent an email with instructions on how to administer MODEST

to their students (see Appendix F5 for the email). Students were given one week to complete the assessment via an online survey platform, Qualtrics. To maximize response rates, the instructors sent a reminder email to their students two days prior to the due date. The useable sample for this cohort was students who consented to participate in the study (see Appendix F6 for the online consent form) and responded to at least one item on MODEST. Based on their responses, MODEST 5 and its test blueprint were modified for a final time (see Appendix A4 for MODEST 6 and see Appendix B3 for its test blueprint). MODEST 6 will be referred to as MODEST for the remainder of this chapter.

3.5 Field Test

The field test phase of this study consisted of administering MODEST to students enrolled in the CATALST course during the Spring 2015 semester. These students completed MODEST twice, once at the beginning of the semester (Pre administration) and once at the end of the semester (Post administration). The rubric was also developed and evaluated during this phase. In this section, the CATALST students and the administration of MODEST to these students are described. In addition, the rubric development process is explained.

3.5.1 Participants. The participants for the field test were students enrolled in EPsy 3264: *Basic and Applied Statistics*–an introductory statistics course at the University of Minnesota–during the Spring 2015 semester. This course, known as the CATALST course, was developed by statistics educators at the University of Minnesota as part of a National Science Foundation grant DUE-0814433 (Garfield, et al., 2012). This introductory statistics course was chosen because it has an explicit learning

objective of "develop students' statistical thinking." To meet this objective, the students learn how to "cook" in statistics rather than just "follow recipes" (Schoenfeld, 1998) through modeling and simulation-based methods, as proposed by Cobb (2007). Students also learn to "cook" (i.e., think statistically) by working cooperatively in groups on activities and using a software tool (i.e., TinkerPlots ™) intended to promote statistical thinking (Garfield, et al., 2012).

This study utilized students from the three face-to-face sections of the CATALST course that were offered during the Spring 2015 semester. The three sections together had approximately 120 students. The students who take this course tend to be undergraduate students majoring in liberal arts programs. They also tend to take the course to fulfill a mathematical thinking course requirement for their major. The three instructors of the CATALST course were graduate students in the Quantitative Methods in Education program at the University of Minnesota.

3.5.2 Data collection. Data collection for the field test occurred twice during the Spring 2015 semester: once during the first week of the semester (Pre administration) and once during finals week (Post administration). Appendix F7 contains the script that was used for recruitment prior to each administration. Following the in-class presentations, the instructors of the CATALST course were sent an email with instructions on administering MODEST to their students (see Appendix F9 for the email). Students were given 4 to 5 days to complete the online assessment during the Pre administration and one week during the Post administration. To maximize response rates, the instructors sent a reminder email to their students one day prior to the due date. For

each administration, the useable sample was students who consented to participate in the study (see Appendix F8 for the online consent form), had only one online submission, and responded to at least one item on MODEST. Then, of the students in the usable samples for each administration, the final sample was students that could be matched between the Pre and Post administrations. The responses from the final sample were used to answer this study's research question.

3.5.3 Rubric development. The process for developing the rubric involved describing in detail how to score the students on each of the elements of statistical thinking in MODEST. To do this, the first step was to describe each element outlining what it means to exhibit that element of statistical thinking. Many of these descriptions came from Wild and Pfannkuch (2002), but additional resources (e.g., MacKay and Oldford, 2000; MacGillivray & Pereira-Mendoza, 2011) were also consulted to better understand some of the elements of statistical thinking (e.g., planning).

Then, after settling on the descriptors, subsets of students' responses from the Pre administration were examined to get a sense of how they demonstrated the intended elements of statistical thinking. Based on this, a decision was made about whether or not each item captured the elements it was intended to assess. After this process, only 16 of the 22 elements of statistical thinking identified in the test blueprint were actually measured in the items composing MODEST.

The next step was to create detailed descriptions for three score categories for each of the 16 elements. The three score categories were "essentially demonstrates" the statistical thinking element (E), "partially demonstrates" the statistical thinking element

(P), or "does not demonstrate" the statistical thinking element (I). To create these descriptions, data from the Pre administration were used. In general, a student's response was categorized as "E" if they demonstrated correct thinking about the element of statistical thinking and elaborated on their thought processes. A student's response was categorized as "P" if they indicated correct thinking about the element, but did not completely elaborate on their thinking or demonstrated inconsistencies in their thinking. A student's response was scored an "I" if it did not meet the criteria for either the "E" or "P" categories.

After the creating the first draft of the rubric, the rubric was refined two more times. First, student responses to the Post administration were examined. The result of this examination was modifying the scoring descriptions to better score responses that were not observed in the Pre administration. The second refinement of the rubric occurred after the rubric was given to an external rater. This rater, a Ph.D. candidate in Statistics Education, used the rubric to score two students' responses. He then provided feedback that resulted in modifying the scoring descriptions and adding another element of statistical thinking to the rubric. Based on this addition, the total number of elements assessed by MODEST increased to 17.

The end result of this process was a rubric that contained descriptions of three score categories ("E", "P", and "I") for each element of statistical thinking being assessed by MODEST. The rubric also included actual student responses for each of the score categories as example responses. To understand the consistency the scoring with the rubric, the inter-rater agreement task is presented next.

3.5.4 Inter-rater agreement.

A second rater, a Ph.D. candidate in Statistics Education, was recruited to analyze a small subset of the students' responses in the field test to assess the consistency of the rubric. This rater was also used in the rubric development phase of the study. To train the second rater on using the rubric, a pilot scoring exercise was carried out on two students' responses in the field test. The second rater and I independently scored these two students' responses and then met to discuss our grading procedures. The rubric was refined to ensure that scoring of the responses was done in the similar manner. Following the pilot scoring exercise, we independently analyzed four students' responses from both the Pre and Post administrations. To maintain consistency in our grading, meetings took place after each scoring of the Pre and Post administrations to compare our scores and discuss our grading procedures. Discrepancies in the scores were discussed and resolved so that 100% agreement was reached.

3.6 Data Analysis of the Field Test

To answer the research question of

What components of students' statistical thinking are revealed and developed in an introductory statistics course that is based on modeling and simulation?,

three "levels" of statistical thinking were examined: results at the element-level of statistical thinking, results at the component-level of statistical thinking, and results at the overall-level of statistical thinking.

To examine the results at the element-level of statistical thinking, the students' score categories ("E's", "P's", and "I's") were compared between the Pre and Post

administration. The results at this level also helped to understand the results at the component-level.

To obtain a single value for each of the four components of statistical thinking, the "E's", "P's", and "I's" were summed across the elements within the components for a student. The E's equaled 1-point, the P's equaled 0.5-points, and the I's equaled 0-points. The totals for each of the components of statistical thinking were

- General Problem-Solving Characteristics = 3,
- Statistical Problem-Solving Processes = 5,
- Cognitive Processes of Statistical Problem-Solving = 6, and
- Individual Dispositions = 3.

Missing responses or responses such as "I don't know" were also coded as 0-points because they lacked information about students' statistical thinking. To answer the research question, confidence intervals using the bootstrap percentile method were computed for each of the four components to understand the development of students' statistical thinking in the CATALST course. In addition, Cohen's *d* was calculated for each component to understand the effect size for each of the components.

A single overall score of statistical thinking was created for both the Pre and the Post administrations of MODEST by using a weighted combination of the four the components of statistical thinking:

$$\textit{Statistical Thinking Score} = w_1*(\textit{General Problem-Solving Characteristics}\ \text{Score}) + w_2*(\textit{Statistical Problem-Solving Behaviors}\ \text{Score})$$

$+ w_3$*(*Cognitive Processes of Statistical Problem-Solving* Score) +

w_4*(*Individual Dispositions* Score).

To ensure each component had equal weight toward the overall score of statistical thinking, the scores for each of the components were converted to a proportion of the component's total score. The overall score of statistical thinking could range from 0 (i.e., demonstrated no statistical thinking on MODEST) to 1 (i.e., demonstrated complete statistical thinking on MODEST).

Due to the complex nature of the assessment, three weighting methods were explored to determine the method that would provide the most meaningful measure of students' statistical thinking. The first method (Method #1) used equal weights (i.e., 0.25) and assumed little to no correlation between the components of statistical thinking. The next two methods used weights generated by the first component from a principal components analysis (PCA). Unlike Method #1, these methods assumed moderate to high correlation between the components. Method #2 used PCA on the scores from the Pre administration to determine the weights for the equation. In contrast, Method #3 used PCA on the scores from the Post administration to find the weights. To decide the appropriate method, the correlation between the components of statistical thinking would be examined.

After applying the chosen method to create the overall score of statistical thinking, confidence intervals using the bootstrap percentile method were computed to also understand the development of students' statistical thinking in the CATALST

course. Then, Cohen's *d* was calculated to understand the effect size for the overall score of statistical thinking.

3.7 Summary of Methods

This chapter described the process of developing, revising, and administering MODEST. This chapter also described the data collection process and analysis used to establish evidence of validity and to answer this study's research question. The results are reported in the next chapter.

Chapter 4

Results

4.1 Introduction

This chapter describes the results of the development and administration of Modeling to Elicit Statistical Thinking (MODEST) assessment. First, three sets of feedback on the assessment are described and the resulting changes to the assessment are summarized. Then, the pilot test administrations of the assessment are presented and changes to the assessment are reported. Finally, the results of the field test administration are described.

4.2 Reviewer Feedback

This section presents the results from three sets of feedback that were solicited to develop and revise MODEST. These sets consisted of feedback from graduate students in the Statistics Education department at the University of Minnesota, expert reviewers who are statistics educators, and a reviewer who edited the assessment. The results from these sets are described and the subsequent changes to MODEST are summarized.

4.2.1 Results of feedback from Statistics Education graduate students.

Three graduate students in the Statistics Education department at the University of Minnesota were asked to complete MODEST 1 from an introductory statistics student's point of view. The primary purpose of this task was to gather preliminary evidence that students' responses to the items were revealing the statistical thinking identified in the test blueprint. The secondary purpose was to gather feedback related to the administration of the assessment. Based on the graduate students' responses, the items seemed to be eliciting the elements of statistical thinking identified in the test blueprint. The graduate

students provided suggestions for improving the administration of MODEST. This included clarifying the description of the assessment task and adding a blank table to analyze the data in the assessment. These suggestions were incorporated into the second version of MODEST (MODEST 2).

4.2.2 Results of feedback from external reviewer. A reviewer with a Ph.D. in Literature was asked edit MODEST 2. Based on her feedback, changes were made to MODEST. These included modifying grammar, revising items to clarify what was being asked, and reorganizing items to ease the task posed in Part II of the assessment. These changes resulted in a third version of MODEST (MODEST 3).

4.2.3 Results of feedback from expert reviewers. Five leaders in the statistics education community were asked to evaluate MODEST as an assessment of statistical thinking, as described by the test blueprint. These expert reviewers consist of statistics educators from different institutions and different countries, as well as a mix of researchers and curriculum developers. The reviewers were asked to give feedback at the item-level and at the assessment-level. The results from their feedback are presented and the subsequent changes to MODEST and the test blueprint are described.

4.2.3.1 Results of feedback from expert reviewers regarding items as measuring an element of statistical thinking. To obtain feedback at the item-level, expert reviewers were asked to indicate the extent to which they agreed or disagreed with the following statement: "The assessment item measures the specified statistical thinking element." Reviewers provided an agreement rating and comments to explain their agreement rating.

Table D1 in Appendix D1 summarizes the agreement ratings of the expert reviewers for each of the items in MODEST 3. The majority of the reviewers selected "Agree" or "Agree, but with Reservations" for all of the elements measured by the items, except for two of the elements: the element of *considers variation* in Item 2 and the element of *analyzes the data* in Item 6. For Item 2, the majority of the reviewers disagreed that Item 2 measured the element of *considers variation*. For Item 6, the reviewers were divided in agreement on whether Item 6 measured the element of *analyzes the data*.

The comments provided by the expert reviewers were also examined (see Appendix D2 for the reviewers' comments). Based off of these comments, changes were made to the items in MODEST. Changes included rewriting items or creating new items to better measure the elements of statistical thinking, removing items that were redundant with other items, clarifying tasks in items, and clarifying the information in the section descriptions. Details of these changes are described in the paragraphs that follow.

Six items were rewritten to better measure the elements of statistical thinking intended in the item. An example of one of these items was Item 2. As previously mentioned, the majority of the reviewers marked "Disagree" that the item measured its intended element of statistical thinking. The reviewers' comments revealed that they felt this item did not capture the key ideas of considering variation (e.g., measurement error, between groups variation, within subjects variation, and sampling variability). Item 2 was rewritten to try to elicit student's understanding of between-subjects variation (see Table 7). Item 6 was also revised to better capture the element of *analyzes the data*. One

reviewer who marked “Disagree” for this item noted that the rank-ordering task in the item seemed to be busywork. Therefore, the rank-ordering task was removed because it did not seem relevant to measuring the element of *analyzes the data*. Other item revisions were made because the reviewers commented that the items prompted the students to directly consider the elements rather than seeing if students would do this naturally. For example, Item 12 attempted to measure the element of *is curious and aware*. The item prompted the students to be curious about a specific aspect of the problem rather than seeing if students were naturally curious. Thus, this item was rewritten to try to assess students’ natural curiosity as they completed the assessment.

Table 7

Example of Rewritten Item

Old Item 2 in MODEST 3
Statistical thinking element: *Considers variation*
3. Are the characteristics you listed of equal value with respect to helping a student be an effective studier? Explain your reasoning.
New Item 2 in MODEST 4
Statistical thinking element: *Considers variation*
1. Do you think that the effect of the study habit factors on course grade would be the same for all students in the course? Explain why the factors would or would not have the same effect on course grade for all students.

Three new items were created to better measure the elements of statistical thinking. Two of the new items were written to measure the element of *reasons with*

statistical models. One of the reviewers commented that the ideas of statistical variability and statistical inference were absent in the assessment. He believed that thinking about the omnipresence of variability is what distinguishes statistical thinking from other kinds of higher-order thinking, such as critical thinking. Based on his comment, these two new items were written to capture students' knowledge of methods of statistical inference.

Two items were removed from the assessment. These items were removed because they were measured the same elements of statistical thinking as those measured in revised or newly created items. For example, Item 9 was trying to measure students' reasoning with statistical graphs for summarizing data. However, a new item in MODEST 4 (i.e., Item 8) was created to elicit similar thinking from students. Therefore, Item 9 from MODEST 3 was deleted from the subsequent versions of MODEST. Appendix A3 includes the fourth version of MODEST, MODEST 4, and shows the resulting changes.

The reviewers' comments were also used to make changes to the test blueprint. One of the changes was revising the descriptions of the elements of statistical thinking to better capture what was being measured in the items. For example, the element of *develops a plan* was modified to include both "data collection" and "data analysis" in the description. This change was made to better capture what was being assessed in Items 1 and 3. Another change was integrating similar elements of statistical thinking. Two elements in the test blueprint–*transforms the raw data into an aggregate form* and *reasons with statistical models*–appeared to be measuring similar thought processes. In fact, the element of *transforms the raw data into an aggregate form* appeared to be a

subset of the element of *reasons with statistical models*. Because of this, the element of *transforms the raw data into an aggregate form* was deleted from the test blueprint and integrated into the element of *reasons with statistical models.* Appendix B2 includes the updated test blueprint for MODEST 4.

4.2.3.2 Results of feedback from expert reviewers regarding MODEST as an assessment of statistical thinking. At the end of the review form, expert reviewers were asked to indicate the extent to which they agreed or disagreed with the following statement: "Overall, the assessment appears to measure statistical thinking, based off of the test blueprint." Reviewers marked their agreement rating and provided comments to explain their agreement choice (see Appendix D3 for the rating results and comments). The results to the agreement ratings are presented in Table 8.

Table 8

Frequency of Reviewers' Agreement Ratings Evaluating MODEST 3 as an Assessment of Statistical Thinking

Agree	Agree, but with Reservations	Disagree
1	3	1

The majority of the reviewers agreed that MODEST appeared to measure statistical thinking. However, many of them had reservations. The reviewers' comments regarding their agreement rating provided suggestions for improving MODEST as an assessment of statistical thinking. These included less hand-holding within the items, emphasizing the notion of *considers variation* throughout the assessment, and reconsidering the elements that are being assessed by each item. These suggestions were

incorporated into the fourth version of the assessment, MODEST 4 (see Appendix A3) and the test blueprint for MODEST 4 (see Appendix B2).

Not all of the reviewers' suggestions were incorporated into the assessment nor into the test blueprint. One reason for not using all of their suggestions was that they were not consistent across the reviewers. For example, one reviewer would suggest adding an element to an item but no other reviewer make that suggestion. Another reason was some suggestions were not detailed enough to be able to make a change. For example, one reviewer did not agree that MODEST measured statistical thinking and commented that she disagreed with the questions rather than the claims of the skills measured by a question. For this reviewer, she was emailed again and asked to clarify her comment but did not respond back.

4.3 Results from Pilot Test

This section presents the results from two cohorts of students that were used for the pilot test of MODEST. The first cohort, senior undergraduate students who were majoring in statistics, was administered MODEST 4 via cognitive interviews. The second cohort was students enrolled in the CATALST course in Fall 2014 semester. They completed MODEST 4 as an online assessment. The results from both of these cohorts are reported and the subsequent changes to MODEST are summarized.

4.3.1 Results from pilot test: Senior statistics students.

Four senior undergraduate students who were majoring in statistics provided responses to MODEST 4 via cognitive interviews. The senior statistics students who participated in the cognitive interviews varied in statistical ability and knowledge; two appeared to be very capable in

applying their statistical knowledge to a new, ill-structured problem and two appeared to be less capable. Three of the students were males and three of them were domestic students. All of the students completed the assessment in an hour and a half or less. However, the time spent on each item varied between the students, as evidence by how quickly the students responded to the items. Two students supplied only their first answer that came to their mind whereas the other two students spent time thinking of the best possible answer.

Observations made during the interview process resulted in modifying MODEST between each of the interviews. This involved modifying how the data in the assessment were presented. The students appeared to struggle with making sense of the data and using the data in the assessment. As a result, the presentation of the data in the assessment was modified throughout the interview process to try to ease the cognitive burden. Table 9 presents the progression of the data changes between each of the interviews.

Table 9

Progression of Changes to the Data in MODEST 4

Modification ID	Description
Original	Qualitative data of fictitious students' responses to *Study Effectiveness Survey* are presented in Part II of the assessment.
1	Qualitative data of fictitious students' responses to *Study Effectiveness Survey* are moved out of the assessment and as an attachment to the assessment.
2	Qualitative and quantitative data of fictitious students' response to *Study Effectiveness Survey* are supplied.
3	Quantitative ratings of the *Study Effectiveness Survey* questions are all placed in the same direction, where high quantitative rating values

indicate better study effectiveness.

In the end, data modifications 1 and 2 were incorporated into the assessment (see Table 9 for data modification IDs). Data modification 3 was not incorporated into the assessment because the student who was presented with data modification 3 disregarded the qualitative meanings of the quantitative ratings. He did this by reverse-coding those ratings that were already reverse-coded. Thus, the data in the assessment had some quantitative ratings that needed to be reverse-coded.

Another change to MODEST during the interview process was the elimination of an item. Item 13 was removed after the first student interview because the content in the item (repeated measurements) is not typically taught in an introductory statistics course. Additionally, the removal of this item made the assessment shorter, which could reduce test burnout.

Following all of the student interviews, the students' responses to MODEST were examined to make additional changes to MODEST. The change was revising four items to better clarify the tasks in the items. For example, Item 4 was originally written to have students think about how survey questions relate to the construct of study effectiveness. However, after reading Item 4, several students attempted to fill out the survey about their own study habits. Item 4 was modified to better clarify its task. Another item that was revised was Item 5. Students either did not understand what was being asked in Item 5 or did not use their responses from Item 4 to answer the question posed in Item 5. Thus, Item 5 was also revised to better clarify its task. See Table 10 for a comparison of Items 4

and 5 between MODEST 4 and MODEST 5. These changes, item modifications and data presentation modifications, resulted in the fifth version of MODEST (MODEST 5).

Table 10

Example of Revised Items

Old Item 4 in MODEST 4
4. To aid in the process of developing an overall score of study effectiveness, fill in the table below describing how you will use each question in creating an overall score. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the score.
Old Item 5 in MODEST 4
5. Report how to produce your overall score of study effectiveness given any student's responses to the survey. Be sure to include any formulas, procedures, or rules on which your score of study effectiveness is based.
Revised Item 4 in MODEST 5
4. When developing an overall score of study effectiveness, you will need to decide how each question on the *Study Effectiveness Survey* will or will not contribute to the overall score. Use the list of questions in the table below to describe how each question will or will not contribute to the overall score of study effectiveness. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the overall score.
Revised Item 5 in MODEST 5
5. Using the answers you gave in question 4, describe how to compute an overall score of study effectiveness for a student.

4.3.2 Results from pilot test: CATALST students.

Students from an undergraduate introductory statistics course, using the CATALST curriculum (Garfield, et al., 2012) were asked to voluntarily complete MODEST 5 via an online survey platform, Qualtrics, at the end of the Fall 2014 semester. As a result, 36 students made up

the usable sample for this cohort (i.e., consented and responded to at least one item). After examining their responses, three changes were made to MODEST. One change was adding an item to help understand the students' methods for analyzing the data in the assessment. Several students appeared to have good methods for analyzing the data in the assessment but were not able to articulate their method in words or symbols. This new item was written as a scaffold to help students describe how they applied their method to data.

Items 10 and 11 were also modified to better elicit their intended elements of statistical thinking. For example, Item 10 was trying to measure the element of *seeks alternative explanations*. It asked students to provide suggestions for revising the survey in the assessment. However, rather than provide suggestions for the survey, students provided suggestions for revising their method for analyzing the data in the assessment. Item 10 was modified to account for responses that provided suggestions for revising either the survey or the method for analyzing the data.

Finally, elements of statistical thinking were modified in three items. For two of the items, they did not appear to be eliciting its intended element of statistical thinking. For example, Item 4 was trying to measure the element of *considers variation*. This element was removed from this item because the question was not eliciting responses related to considering variation. For the third item, Item 13, an element of statistical thinking was added to the item. This change occurred because several students provided responses that only included anecdotal evidence rather than indicating the need to collect data. Therefore, the element of *recognizes the need for data* was added to Item 13. This

element was originally not included in the blueprint for MODEST. However, as a result of adding Item 13 during the assessment development process and the responses gathered from the pilot CATALST students, it appeared that the element of *recognizing the need for data* was measured in MODEST.

The updated assessment, MODEST 6, is in Appendix A4 and its test blueprint is in Appendix B3. The assessment also shows the resulting changes from MODEST 5 to MODEST 6. The end result of this assessment development process was an assessment that contained 13 items and measured 22 elements, with 14 of the 22 elements being unique. MODEST 6 will be referred to as MODEST for the remainder of this chapter.

4.4 Results from Field Test

This section presents the results from the field test. The field test consisted of administering MODEST to students enrolled in the CATALST course during the Spring 2015 semester. These students completed MODEST twice, once at the beginning of the semester (Pre administration) and once at the end of the semester (Post administration). As a result, 108 students made up the useable sample (i.e., consented, had only one online submission, and responded to at least one item) for the Pre administration and 97 students made up the usable sample in the Post administration. Then, 88 of the students in the usable samples could be matched between the Pre and Post administrations. These 88 were the final sample that was used for the analysis of the field test results.

This section is organized in the following manner. The modifications to the elements of statistical thinking during the rubric development process are described first.

This is followed by the results of the inter-rater agreement task. Then, the results of the CATALST students' statistical thinking are reported.

4.5.1 Summary of modifications to the elements of statistical thinking.

Appendix E2 contains a table that displays the changes made to the elements of statistical thinking as a result of creating the rubric. It also has the elements' descriptions as presented in the rubric. In sum, the modifications included dropping elements from an item and revising descriptions of the elements. Details of these modifications are summarized next.

Seven items dropped their intended element of statistical thinking. Elements were dropped from four items–Items 1, 9, 10, and 13–because after examining student responses, it appeared that there was overlap in descriptions of elements within the same item. For example, Item 10 dropped the element of *integrates the statistical and contextual information* because this element was being assessed by the element of *draws a conclusion* in that same item. Elements were also dropped from two items because the items did not appear to elicit its intended elements of statistical thinking in the students' responses. For example, Item 11 originally tried to assess the element of *seeking alternative explanations*. However, the students did not appear to seek alternative explanations as they responded to the item. They, on the other hand, appeared to be thinking critically. Therefore, Item 11 dropped the element of *seeking alternative explanations* and added the element of *is skeptical/critical*. Finally, one item, Item 6, dropped its element of *analyzes data*. This item was added to MODEST 6 to help better assess students' thinking for Items 5 or 7. Additionally, the responses to this item did not

contribute to understanding students' ability to analyze data. For this reason, it was decided that Item 6 would not assess any elements.

Descriptions for two elements of statistical thinking were also revised during the rubric development process. The description for the element of *creates a model* was originally too vague. Consequently, it needed to be clarified to better assess it in Items 4 and 5. According to MacKay (n.d.), "development of conceptual models is first step in developing more detailed quantitative models" ("Why Use Conceptual Models", para. 2). He recommends developing a "conceptual framework for understanding before introducing equations" ("Why Use Conceptual Models, para. 6). Based on this information, creating a model appears to involve both a conceptual model and a quantitative model. Therefore, the element of *creates a model* was split into two sub-elements: *produces a conceptual model* for Item 4 and *translates a conceptual model into a statistical model* for Item 5. Another sub-element, *produces a quality model* for Item 5, was added. This addition occurred because several students adequately described a model but their model did not take into account the nuances of the survey and the data in the assessment. These additions to the element of *creates a model* portray a more complete picture of how students are thinking statistically.

The other element whose description was modified was the element of *draws a conclusion* in Item 10. The description for *draws a conclusion* included the aspects of interpretation, communication, and evaluation, which made the description too broad. Furthermore, the aspects of communication and evaluation in *draws a conclusion* overlapped with other elements that were being assessed in Item 10. As a result, it was

decided that only the aspect of interpretation in the *draws a conclusion* element would be assessed in Item 10.

Appendix E1 contains the final rubric used to score the student responses to MODEST and Appendix B4 compares the original and final test blueprint for MODEST.

4.5.2 Results for inter-rater agreement. A PhD candidate in Statistics Education along with this research provided data to evaluate the consistency of the rubric. This task included scoring a small sample of student responses to the Pre and Post administrations of MODEST. The rubric was applied to four students across 16 elements of statistical thinking in the Pre administration and 17 elements in the Post administration, for a total of 64 individual scores for the Pre administration and 68 individual scores for the Post administration. The overall percent of agreement for the responses scored in the Pre administration was 72% (46 out of 64) with a kappa coefficient of 0.56 (95% CI: 0.40 to 0.72). The overall percent agreement for the responses scored in the Post administration was 63% (43 out of 68) with a kappa coefficient of 0.39 (95% CI: 0.21 to 0.57). For conflicts in the scoring of both administrations, I scored more responses lower than the second rater (14 responses out of the 18 conflicts in the Pre administration data and 18 responses out of the 25 conflicts in the Post administration data). Four scores differed by more than 2 score categories ("E" verses "I") on the responses to both the Pre and Post administrations, for a total of eight total. The responses that had score disagreements were discussed until an agreed upon score was reached.

Table 11 displays the inter-rater percent of agreement of the scoring task. Two elements had very low percent of agreement in the scoring of the Pre administration data: *draws a conclusion* and *is skeptical/critical*. Both of these elements were assessed in Item 10. Twelve of the 16 elements had 75% or above inter-rater agreement in the scoring of the Pre administration data. Five of these had perfect agreement.

Table 11

Inter-Rater Percent of Agreement by Element for each Item

Item	Element	Agreement (Pre)	Agreement (Post)
1	*Develops a [reasonable] plan for collection or analysis of the data.*	75%	75%
2	*Considers variation*	75%	50%
3	*Develops a [reasonable] plan for collection or analysis of the data.*	50%	50%
4	*Creates model: Produces a conceptual model*.	100%	100%
5	Element 1: *Creates a model: Translates the conceptual model into a statistical model*	75%	75%
	Element 2: *Creates a model: Produces a quality model.*	75%	50%
7	*Analyzes the data*	100%	75%
8	*Considers variation*	75%	75%
9	*Reasons with statistical models*	100%	75%
10	Element 1: *Draws a conclusion*	25%	50%
	Element 2: *Reasons with statistical models*	-	25%
	Element 3: *Is skeptical/critical*	0%	25%
11	*Is skeptical/critical*	75%	100%
12	*Is curious*	50%	75%
13	Element 1: *Develops a plan for collection or analysis of the data.*	75%	75%
	Element 2: *Reasons with statistical models*	100%	50%
	Element 3: *Recognizes the need for data*	100%	50%

For the scoring of the Post administration data, two elements had very low inter-rater percent of agreement: *reasons with statistical models* and *is skeptical/critical*. Similar to the scoring of the Pre administration data, both of these elements were assessed in Item 10. Nine of the 17 elements had 75% or above inter-rater agreement inter-rater agreement in the scoring of the Post administration data, with two having perfect agreement.

4.5.3 Results of field test in assessing students' statistical thinking. The final rubric was applied to the 88 CATALST student responses for the Pre and Post administrations. To answer the research question of

What components of students' statistical thinking are revealed and developed in an introductory statistics course that is based on modeling and simulation?,

the results of the students' scores for three levels of statistical are presented.
The three levels, which are hierarchical, are: elements of statistical thinking, components of statistical thinking, and the overall construct of statistical thinking.

First, the results for the elements of statistical thinking are summarized within their respective component. These are presented in two formats. The first format is alluvial plots. These plots visually display the change in students' scores at the element-level between the Pre and Post administrations. They also display the marginal percent of students' scores at each administration. The size of the white blocks on the edges of the plot and the size of the stream fields in the middle represent the percent in each score category. The colors in the stream field can be interpreted as follows:

- Green represents the students who increased their scores between the Pre and Post administrations,
- Purple represents the students who decreased their scores between the Pre and Post administrations, and
- Tan represents the students who remained the same between the Pre and Post administrations.

A table of students' score movements from Pre to Post administration is the second format. The score movements are categorized as an increase in a student's score (i.e., "I" to "P" or "P" to "E"), a decrease in student's score (i.e., "E" to "P" or "P" to "I"), or a score that remained the same from the Pre to Post administration.

Then, the results for the four components of statistical thinking are presented descriptively and inferentially. These four components were created by summing the score categories (i.e., E's, P's, and I's) across their elements of statistical thinking. The "E" category equaled 1-point, the "P" category 0.5-points, and the "I" category 0-points. The totals for each of the components of statistical thinking were

- General Problem-Solving Characteristics = 3,
- Statistical Problem-Solving Processes = 5,
- Cognitive Processes of Statistical Problem-Solving = 6, and
- Individual Dispositions = 3.

The results for the students' overall score of statistical thinking are also presented. Similar to the results at the component-level, these results are presented descriptively and inferentially.

4.5.3.1 Summary of time to completion. To get a sense of how long it took students to complete MODEST, the time to completion was examined. The majority of students completed MODEST in one sitting (i.e., four hours or less); 69 out of 88 students (78%) in the Pre administration and 75 out of 88 (85%) in the Post administration. Of those that completed MODEST in one sitting, the average time to completion was approximately 90 minutes (SD = 52) and 80 minutes (SD = 41) for the Pre and Post administrations, respectively.

4.5.3.2 General Problem-Solving Characteristics component. The first component that is reported is the General Problem-Solving Characteristics component. The results of the elements that made up this component are first presented, followed by the results at the component-level.

4.5.3.2.1 Results for the elements in the General Problem-Solving Characteristics component. The three elements of statistical thinking that made up the component of General Problem-Solving Characteristics were

- *Produces a conceptual model*,
- *Translates the conceptual model into a statistical model*, and
- *Produces a quality model*.

One item assessed the element of *produces a conceptual model* and one item assessed the elements of *translates the conceptual model into a statistical model* and *produces a quality model*. Figures 2 to 4 display the results of these elements.

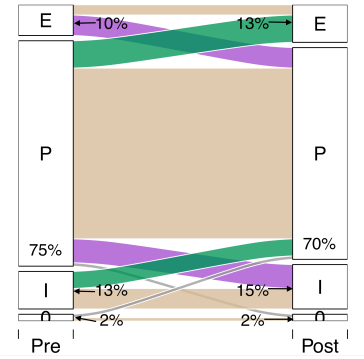

Figure 2. Alluvial plot for the element of *produces a conceptual model* (Item 4).

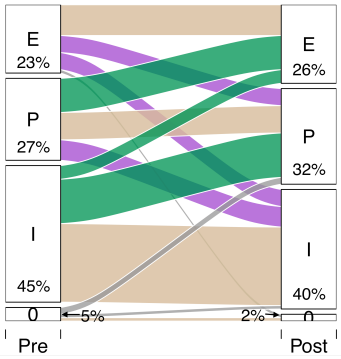

Figure 3. Alluvial plot for the element of *translates the conceptual model into a statistical model* (Item 5).

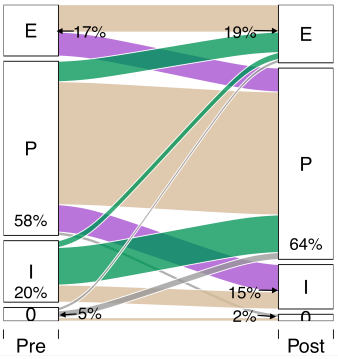

Figure 4. Alluvial plot for the element of *produces a quality model* (Item 5).

The majority of students partially demonstrated the elements of *produces a conceptual model* and *produces a quality model* in both the Pre and Post administrations (see Figures 2 and 4). For the element of *translates the conceptual model into a statistical model*, more students were not able to demonstrate this element in both the Pre and Post administrations than the other score categories (see Figure 3). For the three elements of statistical thinking in this component, there was not meaningful change in the score category percents from the Pre to Post administration. The changes in the percents ranged from a decrease of 5% (i.e., the "P: category in the element of *produces a quality model* and for the "I" category in the element of *translates the conceptual model into a statistical model*) to an increase of 6% (i.e., the "P" category for the element of *produces a quality model*).

Table 12 presents the students' score movements from the Pre to the Post administration for the elements in the component of General Problem-Solving Characteristics. For the three elements in this component, more students obtained the same score between the two administrations than increased or decreased. The element of

translates the conceptual model into a statistical model had more students increase their scores from the Pre to the Post administration than any of the other elements.

Table 12

Students' Score Movement from the Pre to the Post Administration for the Elements in General Problem-Solving Characteristics Component

Element	Increase	Same	Decrease	Total[a]
Produces a conceptual model (Item 4)	13 (15.1%)	60 (69.8%)	13 (15.1%)	86
Translates the conceptual model into a statistical model (Item 5)	27 (32.1%)	41 (48.8%)	16 (19.0%)	84
Produces a quality model (Item 5)	19 (22.6%)	50 (59.5%)	15 (17.9%)	84

Note. The data are presented as count (percent).
[a]The total count for each element is the number of students who have a score for that element in both the Pre and Post administrations (i.e., no missing responses).

4.5.3.2.2 Results for the General Problem-Solving Characteristics component.

For the component of General Problem-Solving Characteristics, students in the Pre administration were, on average, partially able to demonstrate this component (mean = 1.3 out of 3). The distribution of this component in the Pre administration appears to be approximately symmetric (see Figure 5). In the Post administration, a similar result to the Pre administration was found (see Table 19). However, the distribution in the Post administration appears to shift toward the higher values (see Figure 5). The inferential results suggest that the change in scores between the Pre and the Post administrations for this component was not statistically significant (see Table 19). Additionally, the effect size for this component was 0.14, which suggests a small effect. Although there was not a meaningful change between the two administrations for this component, more students

increased their score from the Pre to the Post administration than those that stayed the same or decreased (see Table 21 and Figure 6).

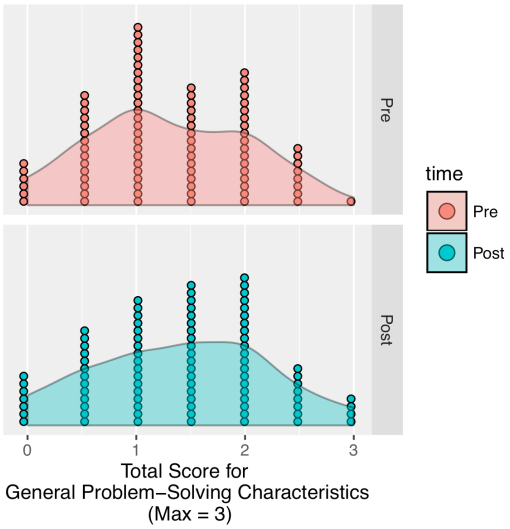

Figure 5. Dotplot, overlaid with density curve, of the scores for the General Problem-Solving Characteristics component by administration.

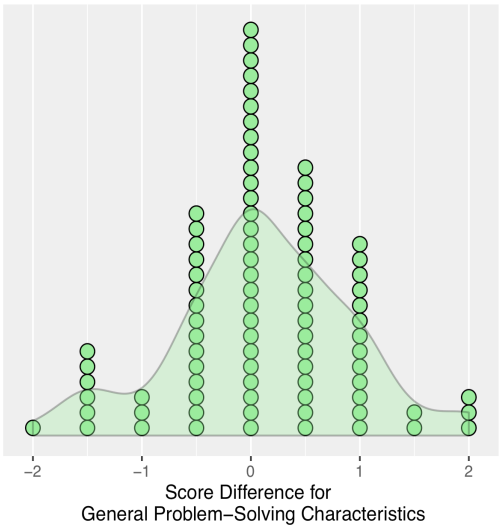

Figure 6. Dotplot, with overlaid density curve, of the score difference for the General Problem-Solving Characteristics component.

4.5.3.3 *Statistical Problem-Solving Processes component.* The next component that is described is the Statistical Problem-Solving Processes component. The results of the elements that made up this component are reported. Then, the results at the component-level follow.

4.5.3.3.1 Results for the elements in the Statistical Problem-Solving Processes component. The three elements of statistical thinking that made up the component of Statistical Problem-Solving Processes were

- *Develops a plan for collection or analysis of the data*,
- *Analyzes the data*, and
- *Draws a conclusion*.

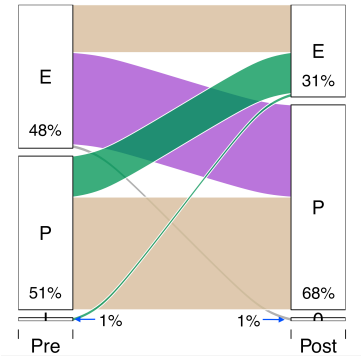

Figure 7. Alluvial plot for the element of *develops a reasonable plan for collection of the data* (Item 1).

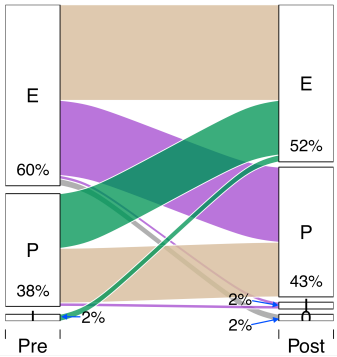

Figure 8. Alluvial plot for the element of *develops a plan for collection of the data* (Item 3).

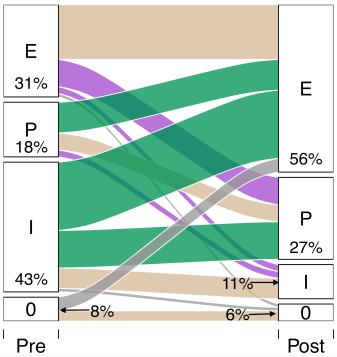

Figure 9. Alluvial plot for the element of *develops a plan for analysis of the data* (Item 13).

Three items assessed the element of *develops a plan for collection or analysis of the data:* two items for *develops a plan for collection of the data* and one item for *develops a plan for analysis of the data*. Figures 7 to 9 display the results of these three items.

In the Pre administration, students were evenly split between essentially and partially demonstrating the element of *developing a reasonable plan for collection of the data* for Item 1 (see Figure 7). However, there were drastic changes in the students' scores between the Pre and Post administrations. In the Post administration, the percent of students who scored an "E" decreased to 30% and the percent of students who scored a "P" increased to 70%. About half of the students obtained the same score from the Pre to the Post administration (54%), while roughly 30% of the students decreased their score from the Pre to the Post administration (see Figure 7 and Table 13).

Item 3 also assessed the element of *develops a plan for collection of the data.* For this item, the majority of students in the Pre administration essentially demonstrated this element (see Figure 8). There were slight changes in the students' scores between the Pre

and Post administration. The percent of students who were scored an “E” decreased by 8% and the percent of students who were scored a “P” increased by 5%. Even with the decrease in the “E” category, a little more than half of the students essentially demonstrated the element of *develops a plan for collection of the data* in the Post administration. About half of the students obtained the same score from the Pre to the Post administration (see Table 13). Furthermore, more students decreased their score than increased (28% verses 21%, respectively).

Figure 9 displays the results the item that assessed the element of *develops a plan for analysis of the data*. Forty-three percent of students in the Pre administration were not able to demonstrate this element, followed by 31% of students essentially demonstrating this element. The comparison of the scores between the two administrations revealed a meaningful increase in the “E” category (i.e., increase by 25%) and meaningful decrease in the “I” category (i.e., decrease by 11%). To further complement this result, 49% of students increased their scores from the Pre to the Post administration (see Table 13).

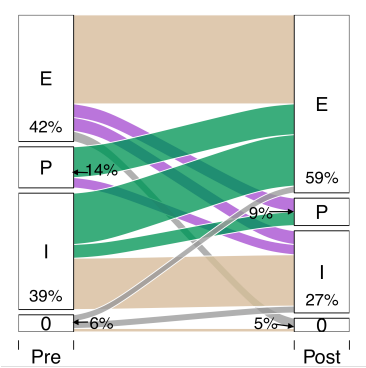

Figure 10. Alluvial plot for the element of *analyzes the data* (Item 7).

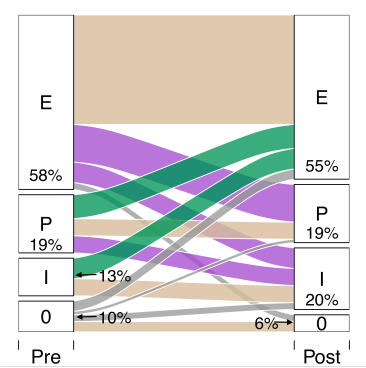

Figure 11. Alluvial plot for the element of *draws a conclusion* (Item 10).

For the element of *analyzes the data*, about 40% of students in the Pre administration essentially demonstrated this element, with the same amount not being able to demonstrate this element (see Figure 10). There was meaningful change between the two administrations for the element of *analyzes the data*. The percent of students who scored an "E" increased by about 20% and the percent of students who scored an "I" decreased by about 10%. Fewer students decreased their scores from the Pre to the Post administration (14%) as compared to those who increased their scores or stayed the same (35% and 52%, respectively) (see Table 13).

For the last element of statistical thinking in this component, *draws a conclusion*, the majority of students in the Pre administration essentially demonstrated this element (see Figure 11). There was little change in the students' scores between the Pre and Post administrations. Furthermore, 56% of the students obtained the same scores in the Pre to Post administrations, followed by about a quarter of the students decreasing their scores from the Pre to Post administrations (see Table 13).

Table 13

Students' Score Movements from the Pre to the Post Administration for the Elements in Statistical Problem-Solving Processes Component

Element	Increase	Same	Decrease	Total[a]
Develops a reasonable plan for collection of the data. (Item 1)	13 (14.9%)	47 (54.0%)	27 (31.0%)	87
Develops a plan for collection of the data (Item 3)	18 (20.9%)	44 (51.2%)	24 (27.9%)	86
Develops a plan for analysis of the data (Item 13)	40 (48.8%)	30 (36.6%)	12 (14.6%)	82
Analyzes the data (Item 7)	28 (34.6%)	42 (51.9%)	11 (13.6%)	81
Draws a conclusion (Item 10)	13 (16.3%)	45 (56.3%)	22 (27.5%)	80

Note. The data are presented as count (percent).

[a]The total count for each element is the number of students who have a score for that element in both the Pre and Post administrations (i.e., no missing responses).

4.5.3.3.2 Results for the Statistical Problem-Solving Processes component. For the component of Statistical Problem-Solving Processes, students in both the Pre and Post administrations were, on average, partially able to demonstrate this component (mean = 3.1 and 3.3 out of 5, respectively). Both of these distributions appear to be left skewed (see Figure 12). Although the change in scores between the Pre and Post administrations for this component was 0.3, this result was statistically significant, albeit small (see Table 19). In line with this result, the effect size for this component (d = 0.28) suggests a moderate effect. About half of the students increased their scores from the Pre to the Post administration for this component (see Table 21 and Figure 13).

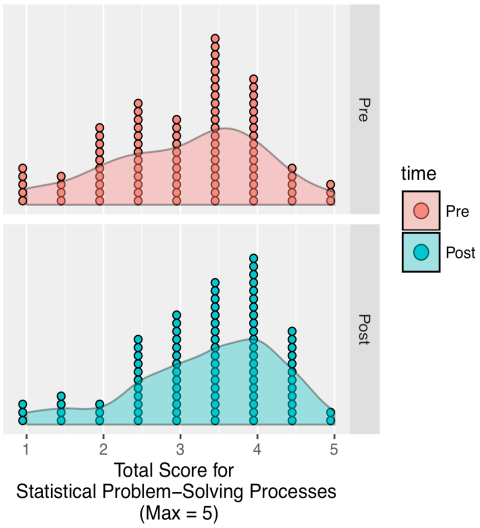

Figure 12. Dotplot, overlaid with density curve, of the scores for the Statistical Problem-Solving Processes component by administration.

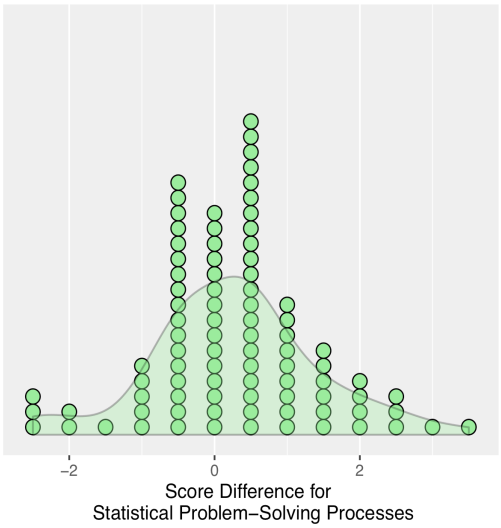

Figure 13. Dotplot, overlaid with density curve, of the score difference for the Statistical Problem-Solving Processes component.

4.5.3.4 *Cognitive Processes of Statistical Problem-Solving component.* The third component is the Statistical Problem-Solving Processes component. The results of the

elements that made up this component and the results at the component-level are now presented.

4.5.3.4.1 Results for the elements of statistical thinking in the Cognitive Processes of Statistical Problem-Solving component. The three elements of statistical thinking that made up the component of Cognitive Processes of Statistical Problem-Solving were

- *Considers variation*,
- *Appropriately reasons with statistical models*, and
- *Recognizes the need for data*.

Two items assessed the element of *considers variation,* three items assessed the element of *appropriately reasons with statistical models,* and one item assessed the element of *recognizes the need for data*.

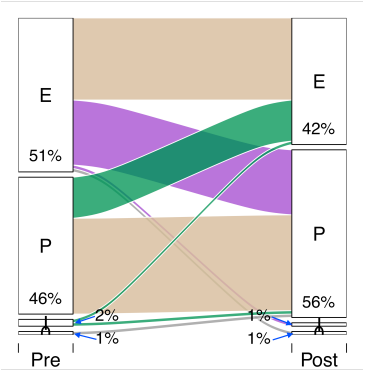

Figure 14. Alluvial plot for the element of *considers variation* (Item 2).

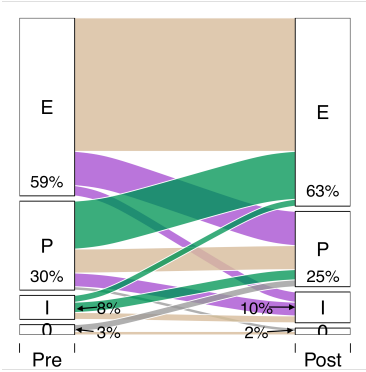

Figure 15. Alluvial plot for the element of *considers variation* (Item 8).

For the two items that assessed the element of *considers variation,* slightly more than half of students essentially demonstrated this element in the Pre administration (see Figures 14 and 15). The change in students' scores between the Pre and Post assessment

varied for this element. In Item 2, the percent of students who scored an "E" decreased by 9% in the Post administration. More students decreased their scores between the Pre and the Post administrations than increased for this item (see Table 14). In contrast, the same element was assessed in Item 8 and students' scores increased slightly in the "E" category in the Post administration. For this item, the percent of students who increased and decreased were roughly the same at 20%.

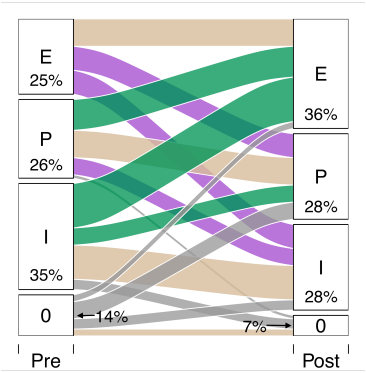

Figure 16. Alluvial plot for the element of *appropriately reasons with statistical models* (Item 9).

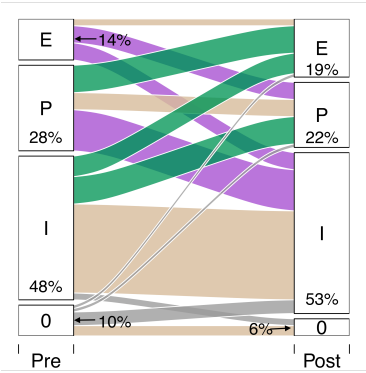

Figure 17. Alluvial plot for the element of *appropriately reasons with statistical models* (Item 10).

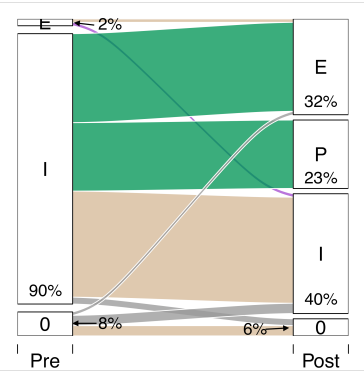

Figure 18. Alluvial plot for the element of *appropriately reasons with statistical models* (Item 13).

Figures 16 to 18 display the results for the three items that assessed the element of *appropriately reasons with statistical models*. Students in the Pre administration were not able to demonstrate this element across all of the items. The percents for the "I" category were the largest among the score categories and ranged from 35% to 90%. The students' scores changed between the Pre and Post administrations for these items. In particular, students' scores in the Post administration increased in the "E" category across all of the items that assessed the element of *appropriately reasons with statistical models*. Item 13

had the largest changes in the score categories. Students' scores increased by 30% in the "E" category and decreased by 50% in the "I" category.

The change within the students' scores varied among the three items for the element of *appropriately reasons with statistical models* (see Table 14). Item 13 had the largest percent of students who increased their scores, at 56%. For Item 10, the same percent of students increased their score as decreased. Item 9 had about the same percent of students who increased their score as those that remained the same.

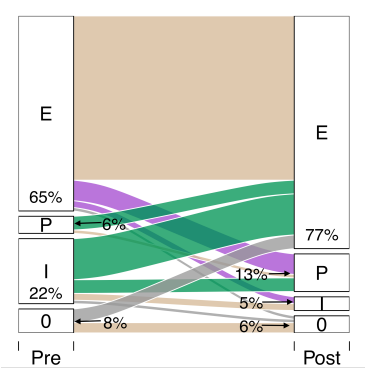

Figure 19. Alluvial plot for the element of *recognizes the need for data* (Item 13).

The last element of statistical thinking in this component is *recognizes the need for data*. The majority of students essentially demonstrated this element in both the Pre and Post administrations (see Figure 19). In addition, there was meaningful change in the students' scores from the Pre to the Post administration. The percent of students who scored an "E" increased by 12% for this element and the percent of students who scored an "I" decreased by 17%. About 66% of students received the same score from the Pre to the Post administration. This was followed by 24% who increased their score and 10% who decreased their score (see Table 14).

Table 14

Students' Score Movements from the Pre to the Post Administration for the Elements in Cognitive Processes of Statistical Problem-Solving Component

Element	Increase	Same	Decrease	Total[a]
Considers variation (Item 2)	14 (16.3%)	52 (60.5%)	20 (23.3%)	86
Considers variation (Item 8)	19 (22.4%)	49 (57.6%)	17 (20.0%)	85
Appropriately reasons with statistical models (Item 9)	27 (36.5%)	28 (37.8%)	19 (25.7%)	74
Appropriately reasons with statistical models (Item 10)	22 (27.5%)	36 (45.0%)	22 (27.5%)	80
Appropriately reasons with statistical models (Item 13)	46 (56.1%)	35 (42.7%)	1 (1.2%)	82
Recognizes the need for data (Item 13)	20 (24.4%)	54 (65.9%)	8 (9.8%)	82

Note. The data are presented as count (percent).
[a]The total count for each element is the number of students who have a score for that element in both the Pre and Post administrations (i.e., no missing responses).

4.5.3.4.3 Results for the Cognitive Processes of Statistical Problem-Solving component. For the component of Cognitive Processes of Statistical Problem-Solving, in the Pre administration, students, on average, were partially able to demonstration this component (mean = 2.8 out of 6) (see Table 19 and Figure 20). Then, in the Post administration, the mean score increased to 3.5 out of 6 possible points for this component. This mean score difference of 0.7 was found to be statistically significant (see Table 19). This component was also found to have a large effect ($d = 0.6$). About 60% of the students increased their score from the Pre to the Post administration for this component (see Table 21 and Figure 21).

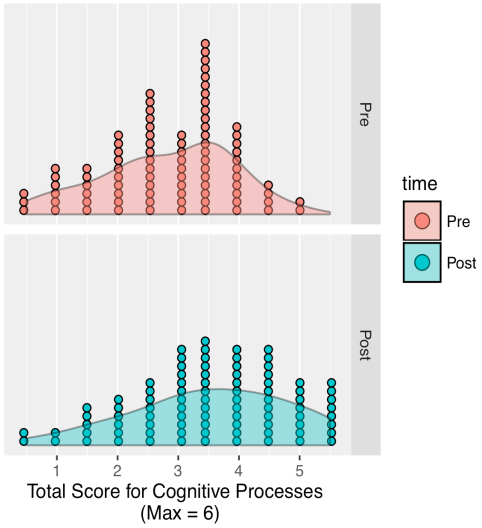

Figure 20. Dotplot, overlaid with density curve, of the scores for the Cognitive Processes of Statistical Problem-Solving component by administration.

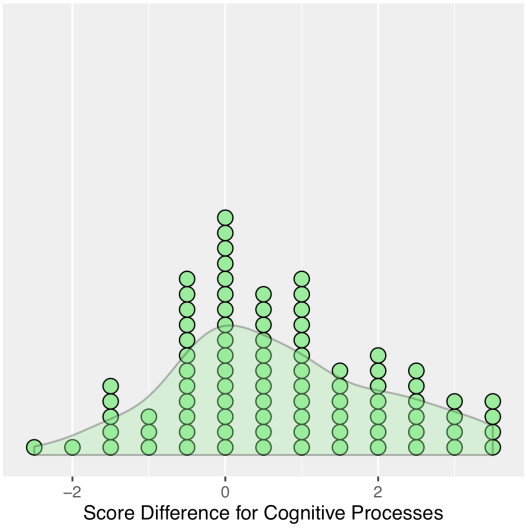

Figure 21. Dotplot, overlaid with density curve, of the score difference for the Cognitive Processes of Statistical Problem-Solving component.

4.5.3.5 ***Individual Dispositions component.*** The last component that comprises statistical thinking is the Individual Dispositions component. The results of the elements for this component are described and then the results at the component-level are summarized.

4.5.3.5.1 Results for the elements of statistical thinking in the Individual Dispositions component. The two elements of statistical thinking made up the component of Individual Dispositions were

- *Is curious* and
- *Is critical*.

One item assessed the element of *is curious* and two items assessed the element of *is critical*. Figures 22 to 24 display the results of these elements.

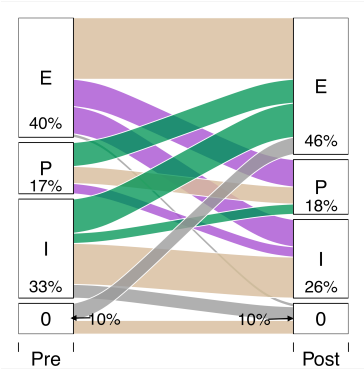

Figure 22. Alluvial plot for the element of *is curious* (Item 12).

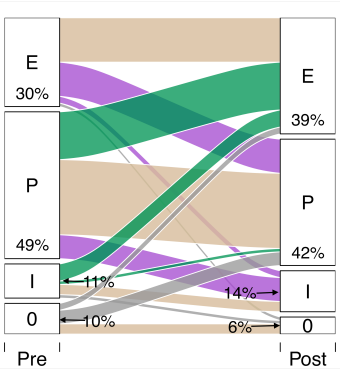

Figure 23. Alluvial plot for the element of *is critical* (Item 10).

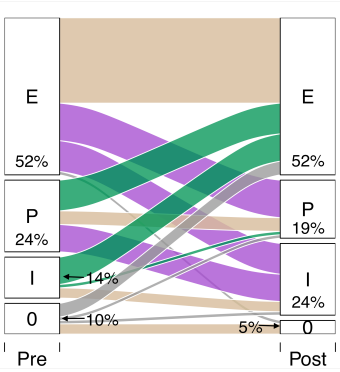

Figure 24. Alluvial plot for the element of *is critical* (Item 11).

Figure 22 displays the results for the element of *is curious*. More students essentially demonstrated this element in the Pre administration than any other score category in the Pre administration. There was little change in the students' score from the Pre to the Post administration. In fact, more students obtained the same score from the Pre to the Post administration than increased or decreased for the element of *is curious* (see Table 15).

The two items that assessed the element of *is critical* reveal different results (see Figures 23 and 24). For Item 10, more students partially demonstrated this element than any other score categories in both the Pre and Post administrations. On the other hand, for Item 11, about half of the students essentially demonstrated the element of *is critical* in both the Pre and Post administrations. Across the two items, more students obtained the same score between the Pre and Post administrations than increased or decreased their scores (see Table 15). For Item 11, more students decreased their scores from the Pre administration to the Post administration than increased their scores. Whereas, for Item

10, roughly the same percent of students increased as decreased their scores from the Pre to the Post administration.

Table 15

Students' Score Movements from the Pre to the Post Administration for the Elements in Individual Dispositions Component

Element	Increase	Same	Decrease	Total[a]
Is curious (Item 12)	20 (25.6%)	39 (50.0%)	19 (24.4%)	78
Is critical (Item 10)	20 (25.0%)	41 (51.3%)	19 (23.8%)	80
Is critical (Item 11)	18 (22.2%)	35 (43.2%)	28 (34.6%)	81

Note. The data are presented as count (percent).
[a]The total count for each element is the number of students who have a score for that element in both the Pre and Post administrations (i.e., no missing responses).

4.5.3.5.2 Results for the Individual Dispositions component. For the component of Individual Dispositions, the students were, on average partially able to demonstrate this component for both the Pre and the Post administrations (mean = 1.7 and 1.8 out of 3, respectively) (see Figure 25). The change in scores was not found to be statistically significant (see Table 19). Furthermore, the effect size for this component was small (d = 0.11). Roughly the same percent of students increased their scores as decreased their scores for this component (42% and 41%, respectively) (see Table 21 and Figure 26).

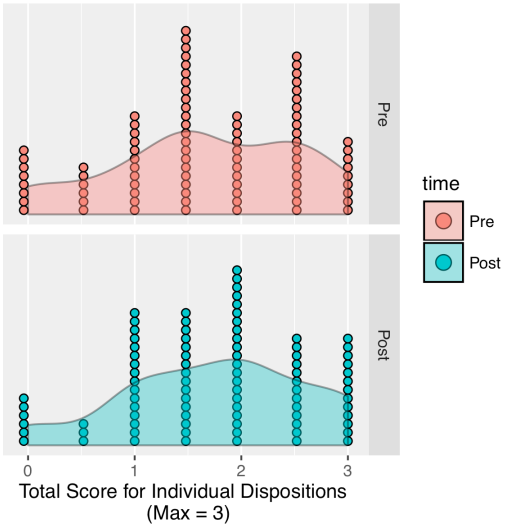

Figure 25. Dotplot, overlaid with density curve, of the scores for the Individual Dispositions component by administration.

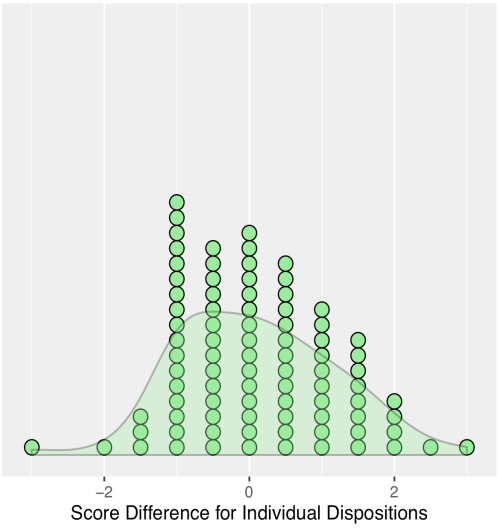

Figure 26. Dotplot, overlaid with density curve, of the score difference for the Individual Dispositions component.

4.5.3.6 Summary of results of field test students' statistical thinking: Element-level.

In sum, there were elements in the Pre administration where the students were able to essentially demonstrate the element of statistical thinking. These elements were

- *recognizes the need for data* (Cognitive Processes in Statistical Problem-Solving component),
- *develops a reasonable plan for the collection of the data* (Item 3; Statistical Problem-Solving Processes component)
- *considers variation* (Item 8; Cognitive Processes in Statistical Problem-Solving component), and
- *draws a conclusion* (Statistical Problem-Solving Processes component).

In contrast, there were elements where the students were not able to demonstrate the elements in the Pre administration (i.e., largest percent in "I" category). These were

- *Appropriately reasons with statistical models* (Item 13; Cognitive Processes in Statistical Problem-Solving component),
- *Appropriately reasons with statistical models* (Item 10; Cognitive Processes in Statistical Problem-Solving component),
- *Translates the conceptual model into a statistical model* (General Problem-Solving Characteristics component),
- *Develops a reasonable plan for the analysis of the data* (Item 13; Statistical Problem-Solving Processes component), and
- *Appropriately reasons with statistical models* (Item 9; Cognitive Processes in Statistical Problem-Solving component).

To understand the development of students' statistical thinking, changes in the students' scores were examined between the Pre and the Post administrations at the element-level. Table 16 displays the nine elements of statistical thinking where little change was seen in the students' scores between the two administrations. Some elements were found to have students essentially demonstrating the element in both the Pre and Post administrations (i.e., *considers variation* (Item 8), *draws a conclusion*). Other elements were found to have students partially demonstrating the element in both the Pre and the Post administrations (i.e., *produces a conceptual model*, *produces a quality model*). Few elements were found to have students not demonstrating the element between the two administrations (i.e., *translates the conceptual model into a statistical model*, *appropriately reasons with statistical models* (Item 10)).

Table 16

Comparison of the Score Distributions for the Elements with Little Change Between the Pre and Post Administrations

Element (Item; Component)	Pre			Post		
	E	P	I	E	P	I
Considers variation (Item 8; CP[c])	59%	30%	6%	63%	25%	10%
Draws a conclusion (Item 10; SPSP[b])	58%	29%	13%	55%	19%	20%
Is critical (Item 11; ID[d])	52%	24%	14%	52%	19%	24%
Is curious (Item 12; ID[d])	40%	17%	33%	46%	18%	26%
Is critical (Item 10; ID[d])	30%	49%	11%	39%	42%	14%
Translates the conceptual model into a statistical model (Item 5; GPSC[a])	23%	27%	45%	26%	32%	40%
Appropriately reasons with statistical models (Item 10; CP[c])	14%	28%	48%	19%	22%	53%
Produces a quality model (Item 5; GPSC[a])	17%	58%	20%	19%	64%	15%
Produces a conceptual model (Item 4; GPSC[a])	10%	75%	13%	13%	70%	15%

Note. The results are sorts in descending order by the "E" category in the Post administration.
[a] GPSC = General Problem-Solving Characteristics component
[b] SPSP = Statistical Problem-Solving Processes component
[c] CP = Cognitive Processes of Statistical Problem-Solving component
[d] ID = Individual Dispositions component

Five elements of statistical thinking were found to have an increase in the students ability to essentially demonstrate the element between the Pre and Post administrations (see Table 17). The difference in the students' scores for the "E" category ranged from 11% to 30%. The largest and smallest increases were seen in the elements of *appropriately reasons with statistical models* (Item 13 and Item 9, respectively). The element of *recognizes the need for data* had a relatively large increase in the "E" category between the two administrations, despite having large percents in the "E" category for the Pre administration.

Table 17

Comparison of the Score Distributions for the Elements with Meaningful Increases Between the Pre and Post Administrations

Element (Item; Component)	Pre			Post		
	E	P	I	E	P	I
Appropriately reasons with statistical models (Item 13; CP[c])	2%	0%	90%	32%	23%	40%
Develops a reasonable plan for the analysis of the data (Item 13; SPSP[b])	31%	18%	43%	56%	27%	11%
Analyzes the data (Item 7; SPSP[b])	42%	14%	39%	59%	9%	27%
Recognizes the need for data (Item 13; CP[c])	65%	6%	22%	77%	13%	5%
Appropriately reasons with statistical models (Item 9; CP[c])	25%	26%	35%	36%	28%	28%

Note. The results are sorts in descending order by the largest increase in the "E" category from the Pre administration to the Post administration.
[a] GPSC = General Problem-Solving Characteristics component
[b] SPSP = Statistical Problem-Solving Processes component
[c] CP = Cognitive Processes of Statistical Problem-Solving component

Three elements of statistical thinking were found to have a decrease in the students' ability to essentially demonstrate the element from the Pre to the Post administration (see Table 18). The largest decrease was 17% in the element of *develops a reasonable plan for the collection of the data* (Item 1; Statistical Problem-Solving Processes component). In addition, the first three items in MODEST measured the three elements with a decrease in the "E" category between the two administrations.

Table 18

Comparison of the Score Distributions for the Elements with Meaningful Decreases Between the Pre and Post Administrations

Element (Item; Component)	Pre			Post		
	E	P	I	E	P	I
Develops a reasonable plan for the collection of the data (Item 1; SPSP[b])	48%	51%	1%	31%	68%	1%
Considers variation (Item 2; CP[c])	51%	46%	2%	42%	56%	1%
Develops a reasonable plan for the collection	60%	38%	2%	52%	43%	2%

of the data (Item 3; SPSP[b])

Note. The results are sorts in ascending order by the largest decrease in the "E" category from the Pre administration to the Post administration.
[b] SPSP = Statistical Problem-Solving Processes component
[c] CP = Cognitive Processes of Statistical Problem-Solving component

4.5.3.6 Summary of results of field test students' statistical thinking:

Component-level. To understand what components of students' statistical thinking were revealed, the scores for the four components in the Pre administration were examined. Across all of the components, the students were partially able to demonstrate the components in the Pre administration. The average percent of possible points ranged from 43% to 62% for the four component scores (see Table 19). The component of General Problem-Solving Characteristics had the lowest average percent in the Pre administration and the component of Statistical Problem-Solving Processes had the highest average percent.

Table 19

Summary Statistics of Scores for the Pre Administration, Post Administration, and Score Difference for Each of the Components

Component	Pre	Post	Difference	95% CI of the Mean Difference	
	Mean (SD)	Mean (SD)	Mean (SD)	Lower Limit	Upper Limit
General Problem-Solving Characteristics (Max Score = 3)	1.3 (0.7)	1.4 (0.8)	0.1 (0.8)	-0.06	0.27
Statistical Problem-Solving Processes (Max Score = 5)	3.1 (1.0)	3.4 (0.9)	0.3 (1.2)	0.03	0.51
Cognitive Processes in Statistical Problem-Solving (Max Score = 6)	2.8 (1.1)	3.5 (1.3)	0.7 (1.4)	0.40	1.00

Individual Dispositions (Max Score = 3)	1.7 (0.9)	1.8 (0.8)	0.1 (1.1)	-0.13	0.32

Note. The difference is calculated as Post – Pre.

Then, to understand what components of students' statistical thinking were developed, the score differences for each component between the two administrations were computed. Two of the four components were found to have a significant increase in average scores from the Pre to the Post administration (see Table 19). These components were Statistical Problem-Solving Processes and Cognitive Processes in Statistical Problem-Solving. The mean score differences for these two components were 0.3 and 0.7, respectively. To further explain these increases, it was found that more of these components' elements had meaningful increases in the students' scores than decreases between the two administrations. The other two components, General Problem-Solving Characteristics and Individual Dispositions, were not found to have a significant change in average scores from the Pre to the Post administration. To further explain the lack of significant change in these components, it was found that these components' elements had little to no change in the students' scores from the Pre to Post administration. For those components with significant increases, they also had moderate to large effect sizes, whereas the components with insignificant changes had small effect sizes (see Table 20).

Table 20

Effect Size Estimates for Each of the Components

Component	Cohen's *d*
General Problem-Solving Characteristics	0.14
Statistical Problem-Solving Processes	0.28
Cognitive Processes in Statistical Problem-Solving	0.58
Individual Dispositions	0.11

Students' score movements were also examined. For the components of Statistical Problem-Solving Processes and Cognitive Processes in Statistical Problem-Solving, more than half of students increased their scores between the Pre and the Post administrations (see Table 21). For the component of General Problem-Solving Characteristics, about 40% of students increased their scores between the Pre and Post administrations. However, for this same component, roughly the same percent of students had the same score as decreased their score between the two administrations (~30% for both). Then, for the component of Individual Dispositions, about the same percent of students increased their score as decreased their score between the Pre and Post administrations (~40% for both).

Table 21

Summary of Score Changes Between the Pre and Post Administrations for Each of the Components

Component	Increase	Same	Decrease
General Problem-Solving Characteristics	36 (41%)	27 (31%)	25 (28%)
Statistical Problem-Solving Processes	45 (51%)	15 (17%)	28 (32%)
Cognitive Processes in Statistical Problem-Solving	50 (57%)	16 (18%)	22 (25%)
Individual Dispositions	37 (42%)	15 (17%)	36 (41%)

Note. The data are presented as count (percent).

4.5.3.7 Results for overall statistical thinking. To create an overall score of statistical thinking, three weighting methods of the four components were explored. The results of this exploration are reported first. Then, the results for the overall score of the students' statistical thinking are presented.

4.5.3.7.1 Results of the exploration of three weighting methods. Three weighting methods for the overall score of statistical thinking were investigated. The purpose of this

investigation was to determine the method that provides the most meaningful measure of students' statistical thinking. The three methods were

- Method #1: equal weight of the linear combination of the components (e.g., 0.25). This method assumes little to no correlation between the components.
- Method #2: weights are generated by the first principal component via Principle Components Analysis (PCA) on the scores from the Pre administration. This method assumes high correlations between the components. The weights from this method would be scaled to equal 1.
- Method #3: weights are generated by the first principal component via PCA on the scores from the Post administration. This method also assumes high correlations between the components. The weights from this method would be scaled to equal 1.

To see whether the components are highly correlated with one another, Spearman correlation coefficients were computed between the four components of statistical thinking (see Table 22). The correlation values range from 0.19 to 0.52 for the scores in the Pre administration and ranged from 0.20 to 0.56 for the scores in the Post administration. Based on these results, it appears that there is low to moderate associations between the four components for each of the administrations.

Table 22

Spearman Correlation Values Between the Four Components of Statistical Thinking

	General Problem-Solving Characteristics	Statistical Problem-Solving Processes	Cognitive Processes of Statistical Problem-	Individual Dispositions (ID)

	(GPSC)	(SPSP)	Solving (CP)	
GPSC	-	0.52	0.19	0.35
SPSP	0.46	-	0.40	0.40
CP	0.40	0.54	-	0.47
ID	0.20	0.44	0.56	-

Note. The correlations for the Pre administration are in the upper triangle and the correlations for the Post administration are in the lower triangle.

Additionally, when the other two methods were carried out, the weights of the linear combination of the components were very close to the weights for Method #1 (i.e., 0.25). The equation produced by the first principle component (PC1) for Method #2 was

$$PC1 = \underline{0.24}\ GPSC + \underline{0.27}\ SPSP + \underline{0.23}\ CP + \underline{0.26}\ ID,$$

with 54.3% of the variance explained by the first principal component. The equation produced by the first principle component (PC1) for Method #3 was

$$PC1 = \underline{0.21}\ GPSC + \underline{0.27}\ SPSP + \underline{0.28}\ CP + \underline{0.24}\ ID,$$

with 58% of the variance explained by the first principal component. Based on the results from the correlation matrix and the PCA analyses from Methods #2 and #3, Method #1 was the chosen method for obtaining the overall score of statistical thinking.

4.5.3.7.2 Results for overall score of statistical thinking. As previously mentioned, the four components had different total scores. Therefore, to ensure each component has equal weight toward the overall score of statistical thinking, the scores for each of the components were converted to a proportion of the component's total score. Then, a student's overall score of statistical thinking was computed by using Method #1:

$$Statistical\ Thinking\ Score = 0.25\ GPSC + 0.25\ SPSP + 0.25\ CP + 0.25\ ID.$$

The overall score of statistical thinking could range from 0 (i.e., demonstrated no statistical thinking on MODEST) to 1 (i.e., demonstrated complete statistical thinking on MODEST).

Table 23 and Figures 27 and 28 display the results of the overall score of statistical thinking for both administrations and comparing the students' score between the two administrations. The average overall score of statistical thinking was 0.52 for the Pre administration was 0.58 for the Post administration. This increase of 0.06 between the two administrations was statistically significant (95% CI: 0.02 to 0.10). The Cohen's effect size value ($d = 0.34$) also indicated a moderate difference between the Pre and Post administrations.

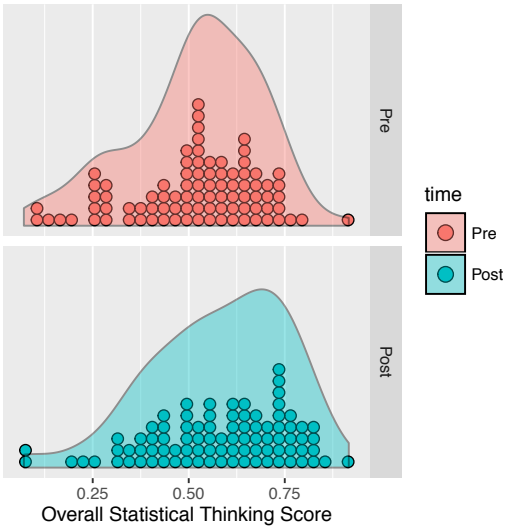

Figure 27. Dotplot, overlaid with density curve, of the overall score of statistical thinking by administration.

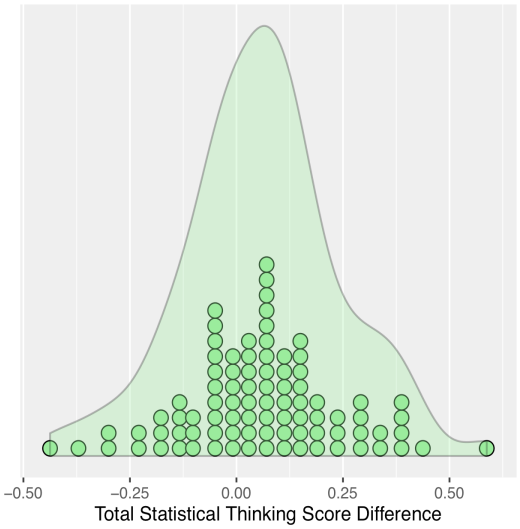

Figure 28. Dotplot, with overlaid density curve, of the overall score difference of statistical thinking.

Table 23

Summary Statistics of Overall Score of Statistical Thinking for Pre Administration, Post Administration, and Score Difference

	Pre	Post	Difference	95% CI of the Mean Difference	
	Mean (SD)	Mean (SD)	Mean (SD)	Lower Limit	Upper Limit
Overall Statistical Thinking Score	0.52 (0.17)	0.58 (0.18)	0.06 (0.18)	0.02	0.10

Note. The difference is calculated as Post – Pre.

Students' score movements of overall statistical thinking were also examined. About 60% of students increased their overall score of statistical thinking between the Pre and Post administrations (see Table 24). The median score difference for those that increased their overall score was 0.13, whereas, the median score difference for those that decreased was 0.10.

Table 24

Summary of Changes in Overall Score of Statistical Thinking Between the Pre and Post Administrations

Increase	Same	Decrease
56 (64%)	1 (1%)	31 (35%)

Note. The data are presented as count (percent). The Pre and Post scores were compared to the nearest hundredth decimal place.

4.6 Summary of Results

This chapter presented the results of the development and administration of MODEST. During the development process, which included soliciting feedback from reviewers and pilot testing MODEST, MODEST was revised to improve MODEST as an assessment of statistical thinking. In the field test phase, CATALST students revealed that they were able to partially think statistically at the beginning of the semester and appeared to increase their ability to think statistically over the semester. The next chapter

summarizes these results, discusses limitations of this study, and provides implications for future research.

Chapter 5

Discussion

5.1 Study Summary

Modeling to Elicit Statistical Thinking (MODEST) was created to assess students' statistical thinking in an introductory statistics course that is based on modeling and simulation. The CATALST course was used for the field test of MODEST and to answer the research question because it has an explicit learning objective of "develop students' statistical thinking." This study consisted of three phases: developing the assessment, pilot testing the assessment, and field testing the assessment, with the goal of answering the research question of

What components of students' statistical thinking are revealed and developed in an introductory statistics course that is based on modeling and simulation?

To develop MODEST, a construct-centered approach to assessment design was used. First, a test blueprint was created to identify characteristics of statistical thinking. Then, items were written around a model-eliciting activity (MEA) context that measured the elements of statistical thinking in the test blueprint. After MODEST was created, several phases of reviewer feedback were collected, including feedback from experts in statistics education, and MODEST was modified after each phase.

Pilot testing followed the assessment development phase. During the pilot test, MODEST was administered to senior undergraduate students who were majoring in statistics via cognitive interviews and to CATALST students via an online assessment. MODEST was again refined after each administration.

Finally, MODEST was given to CATALST students at the beginning and end of their course to understand what statistical thinking is revealed and developed. The students' responses to MODEST were used to create a scoring rubric. After applying the rubric to the responses, three levels of statistical thinking were examined and compared to answer the research question. These levels were element-level, component-level, and overall-construct of statistical thinking.

5.2 Summary of the Results

Prior to this study, there was no known assessment that tried to measure the complete construct of statistical thinking. Therefore, to answer the research question, MODEST was created as an assessment of statistical thinking. MODEST used a type of problem that was suggested in the literature for assessing expert-like thinking (i.e., an MEA) and also used a test blueprint of statistical thinking. Evidence supporting MODEST as an assessment of statistical thinking is first presented and then the field test results are summarized to provide an answer to the research question.

5.2.1 Validity evidence of MODEST.

Guidelines for assessment development were used to try to ensure that MODEST was an assessment of statistical thinking (e.g., American Psychological Association and the National Council on Measurement in Education, 1999). Of importance were the use of the test blueprint to identify the characteristics of statistical thinking that would be measured in MODEST, the expert feedback to evaluate MODEST as an assessment of statistical thinking, and the multiple sources of student responses to help refine MODEST before the field test. Another key

step of the process was the use of a second rater to understand the consistency of the scoring rubric.

The test blueprint in this study provided the first set of evidence for MODEST as an assessment of statistical thinking. This test blueprint of statistical thinking was created and used throughout the study. It incorporated domain-general and domain-specific expert characteristics that were found in the literature to provide a more complete understanding of what it means to think like an expert within the domain of statistics. Furthermore, the items for MODEST were written to measure the elements of statistical thinking that were defined the test blueprint. Then, after each data collection (i.e., reviewer feedback or student responses to MODEST), the test blueprint was revisited and updated to ensure that it was capturing the statistical thinking that was elicited in MODEST. The test blueprint also formed the basis for the scoring rubric. Consequently, by using the test blueprint from start to finish, this provides evidence of MODEST as an assessment of statistical thinking.

Expert feedback to MODEST also provide evidence toward MODEST as an assessment of statistical thinking. The majority of the expert reviewers agreed that the items in MODEST 3 measured their intended elements of statistical thinking, except for two. These two items were revised to better capture their intended elements. The majority of the expert reviewers also agreed that MODEST 3 appeared to measure statistical thinking, as defined by the test blueprint. Even with overall positive nature of their feedback, the experts' ratings and comments contributed greatly to the revision of MODEST by creating or rewriting items to better assess particular elements of statistical

thinking. The test blueprint was also updated to better define the elements of statistical thinking that is captured in MODEST. The most substantial changes made based off of the experts' feedback were revising items to better capture the idea of variation, writing new items to understand students' natural habits of mind, and writing new items on statistical inference. The experts' evaluation and the revisions made based on their feedback provide additional evidence of MODEST as an assessment of statistical thinking.

Gathering data on the consistency in scoring is the final set of evidence for MODEST as an assessment of statistical thinking. To understand this, a second rater was recruited to score a small number of students. Overall, the results of the overall inter-rater agreement were moderately high for the scoring of the Pre administration responses and moderate for the Post administration responses. In addition, the majority of the inter-rater agreements at the element-level were high for both administrations. However, one of the items, Item 10, consistently had low inter-rater agreement results in the scoring of its elements. In particular, the author scored the responses at a lower score category than the second rater for all of the score discrepancies. Because of this, the second rater and the author discussed the descriptions in the scoring rubric for this item. The rubric was updated to provide clarification on how the elements would be scored.

In sum, multiple steps were taken to provide evidence of validity for MODEST as an assessment of statistical thinking. These steps included using a test blueprint from start to finish, soliciting feedback from expert reviewers to make changes to MODEST, and using a second rater to gather data on the consistency in scoring. Taking all of these steps

and their results into consideration, there is evidence that MODEST is an assessment of statistical thinking.

5.2.2 Students' statistical thinking. To answer the research question, the results of the overall score of statistical thinking and the components of statistical thinking were examined. The field test results from the Pre administration were used to understand and reveal students' statistical thinking. Then, the field test results comparing the Pre to the Post administration were used to determine how statistical thinking developed during the course.

In the Pre administration, the field test results revealed that students come into the CATALST course with some ability to think statistically (i.e., think like an expert statistician). That is, students were able to *partially* demonstrate thinking statistically at the beginning of the semester, based on the average overall statistical thinking score of a little more 50%. To understand this result in more detail, results at the component-level were examined. For all of components of statistical thinking, students were able to *partially* demonstrate thinking related to the components in the beginning of the semester. While it is known that students enter a course with pre-existing knowledge (e.g., Hidi & Anderson, 1992), it is interesting that these students are able to demonstrate thinking statistically at a moderate level at the beginning of the course. However, this result may not be surprising, given that many of the components of statistical thinking are not unique to the field of statistics (e.g., individual dispositions, general problem-solving characteristics). These students have most likely used several of these domain-general characteristics of problem-solving in their daily life and in their education. Another

potential reason why students partially demonstrated statistical thinking at the beginning of the semester is that the rubric might not be sensitive enough to tease out the novices from those that have more expert-like thinking.

To examine the development of statistical thinking, the field test results were compared between the Pre and Post administrations. The results indicated that students' overall statistical thinking scores increased between the two administrations. Although this increase at the overall statistical thinking level was significant, the average percent increase was small. Furthermore, for all four of the components, students' scores increased between the Pre and Post administrations, with two of the four components having significant increases (i.e., Statistical Problem-Solving Processes and Cognitive Processes in Statistical Problem-Solving). The lack of a significant change in the component of Individual Dispositions is consistent with results from previous studies (e.g., Pfannkuch and Wild, 2003); that is, students struggle to develop dispositions of thinking statistically. In sum, the results at the overall-level and component-level suggest that students do develop some statistical thinking in the CATALST course. This is consistent with other studies that assess the development of novices' thinking in a course (e.g., Chin & Chia, 2005; Derry, et al., 2000). However, given the small increase in the overall statistical thinking score and the modest increases at the component-level, these data suggest more could be done in the CATALST course to increase students' statistical thinking. Some recommendations for how the CATALST course can better develop students' statistical thinking are in the implications for teaching section of this chapter (see Section 5.4).

An examination of students' responses to the instrument offer some insights into the results. For example, the majority of students were essentially able to demonstrate the element of *recognizes the need for data* at the beginning of the semester, and even had a meaningful increase at the end of the semester. This provides one piece of evidence that students have some expert-like thinking when they enter the course and is in contrast to the result found in Pfannkuch and Wild (2003). In their study, they reported that acknowledging the need for data was a student difficulty. Based on this result, it appears that many students enter and exit the course understanding that data is needed to make reasonable conclusions about a question. It was also observed that at the end of the course, students were better able to explain their thought processes using statistical terminology (e.g., random assignment, confidence intervals, randomization tests) but were not always able to decipher when those methods would be used. For example, when asked to describe how to summarize the data in the assessment, some students described using inferential methods (e.g., bootstrapping a confidence interval, conducting a randomization test) to summarize data from a sample. This result is consistent to what is found in the expert-novice literature, which is novices struggle to adapt and transfer their knowledge to solve new problems (e.g., Bransford et al., 2000). Based on this result, it appears that some CATALST students are unclear whether inferential methods verses descriptive methods would be useful to analyze data. One possible reason for this confusion could be due to the focus on the logic of inference in the course. In the curriculum, the majority of the time is spent on understanding modeling and simulation and little time is spent on using descriptive methods to understand the data.

5.3 Study Limitations

Although MODEST can be considered an assessment of statistical thinking, there are limitations to the conclusions that are made in this study. First, the results can only be generalized to students in the CATALST course at the University of Minnesota. This cohort may not be representative of other institutions that use the CATALST curriculum or other introductory statistics courses that are based on modeling and simulation.

This study is also limited by the characteristics of statistical thinking that are assessed by MODEST. In particular, MODEST measured only a subset of the characteristics of statistical thinking. There were characteristics of statistical thinking that were either left out of the test blueprint (e.g., perseverance, engagement) or were removed from the test blueprint during the development of the rubric due to lacking evidence of the thinking in students' response (e.g., is open to new ideas, seeks alternative explanations). Therefore, the inferences that can be made about students' statistical thinking are limited to those characteristics of statistical thinking in the final test blueprint.

Administration issues of MODEST are of concern, as well. Due to the complex nature of the tasks and the constructed-response type of items in the assessment, students could have taken many hours to adequately complete MODEST. As a result, students could have experienced cognitive burnout. Another issue could be student motivation. The field test students were required to complete MODEST as part of their course grade, but were graded on completion, not correctness. Therefore, the amount of effort students put into completing MODEST could affect the ability to adequately evaluate their

statistical thinking. Another concern is related to completion time. About 20% of students in both administrations completed the assessment over multiple sittings (i.e., more than 4 hours). This is concerning because students could have discussed the assessment with others between the sittings or been distracted while taking the assessment. All of these administration issues could add measurement error to the results.

Although steps were taken to try to better clarify how an item was worded, it appears that there were still interpretation issues in MODEST. For example, many students equated the phrase "study effectiveness" with "getting a good grade." While in practice, the two may be highly correlated, the goal of the problem in MODEST was to evaluate fake students' study habits and see if their habits were effective rather than evaluate whether the fake students received a good grade. Therefore, using the word "effectiveness" in the assessment could be problematic. Another example is using the word "summary" in Item 10. Students seemed to describe or state the results from their analysis rather than summarize them in a statistical manner. Due to these interpretation errors, measurement error could be introduced to the results.

As discussed previously, there was evidence that students increased their statistical thinking in the CATALST course, but the increase was modest. Potential factors that could have affected the scores could be the students, the curriculum, the scoring of the assessment, or the instruction of the material. However, because this was an observational study, it is hard to know what factors contributed to the small (and, for some components, insignificant) change in statistical thinking.

5.4 Implications for Teaching

As recommended by the GAISE College Report (American Statistical Association, 2016), introductory statistics courses should have a learning objective of "teach statistical thinking." To this end, MODEST can be a useful instrument for assessing the development of students' statistical thinking. Instructors could use the results from MODEST to understand what statistical thinking is present in students prior to instruction. This information could inform their teaching by helping them identify the components of statistical thinking that are initially lacking. Then, although this study found a small increase in students' statistical thinking at the overall-level and component-level, instructors could compare of the results between the Pre and Post administrations at the element-level to understand what statistical thinking develops during the semester. This information can inform instructors on what characteristics of statistical thinking students struggle with and therefore can inform their teaching for future semesters. Students could also use the results at the element-level to understand their strengths and weaknesses of statistical thinking.

For the course that was used in this study, the CATALST instructors could use the results from MODEST to better develop students' statistical thinking. Two components did not have significant improvement in the CATALST students' scores: Individual Dispositions and General Problem-Solving Characteristics. Specifically, CATALST instructors could better integrate *being curious*, *being critical* about information, and *creating a model* in the course. Furthermore, more could be done in the course to see a larger improvement in the two other components of statistical thinking (i.e., Cognitive Processes of Statistical Problem-Solving and Statistical Problem-Solving Processes).

Elements of statistical thinking that could be targeted include *reasoning about statistical models*, especially descriptive or graphical summaries, and *developing a plan for collection and analysis of data*. To develop all of these components, the literature suggested using question prompts (e.g., procedural prompts, elaboration prompts, and reflection prompts), expert modeling, and ill-structured problems to develop expert-like thinking.

5.5 Implications for Future Research

This study was an exploratory study on developing an assessment of statistical thinking and assessing students' statistical thinking. Because of this, there are many opportunities for future research. One area of future research is to further refine MODEST. As previously mentioned, the statistical thinking that was measured by MODEST is only a subset of the characteristics of statistical thinking. Therefore, items could be revised or added to assess the missing characteristics of statistical thinking (e.g., engagement, seeking alternative explanations). Items or the scoring of the items could also be revised to better discriminate students' ability to think statistically. There were a few items that appeared to not have good discrimination. In Item 1, for example, the purpose was to force students to grapple with the problem in MODEST (i.e., study effectiveness) and assessed the element of *develops a plan for collection of the data*. However, it appeared that the majority of the students could develop a plan, as asked in the item's stem. This result could be due to the item's task requiring lower cognitive ability for developing a plan or due to the scoring of the item. Thus, to better differentiate

students' statistical thinking, items or their scoring could be modified to elicit or detect a wider range of student responses.

Another area of research could investigate whether different contexts better elicit and assess students' statistical thinking. The context of MODEST used an MEA as the foundation for the ill-structured problem. The primary task in this MEA, called the Study Effectiveness MEA, was to have students create a summary score of study effectiveness for a survey about study effectiveness. Because of this, the majority of the assessment was centered on descriptive methods and only one question required students to think about inferential methods. Therefore, to better target the whole investigative process of solving a statistical problem (e.g., research question, problem representation, data collection, data analysis, conclusions), other ill-structured statistical problems could be considered for an assessment of statistical thinking.

Additional research could also consider administering MODEST to other cohorts of statistical thinkers. Results from students in other introductory statistics courses could be compared to understand the effect of a course on students' statistical thinking. For example, the three known introductory statistics curricula based on modeling and simulation–the CATALST course (Garfield, et al., 2012), the Lock textbook (Lock, et al., 2013), and the ISI textbook (Tintle, et al., 2013)–could be compared to understand how the curricula impact students' statistical thinking. In addition, in light of the current focus on undergraduate statistics programs across the U.S., it would be useful to administer MODEST to senior statistics majors. While a small sample of data from senior statistics majors were gathered in this study, there is preliminary evidence that senior statistics

majors vary widely in their ability to think statistically. Results from a larger sample could inform statistics programs on their ability to develop their students' statistical thinking. Another implication for future research could be administering MODEST to statisticians to understand what statistical thinking occurs and is revealed by those who are experts. Then, the results from the three cohorts could be compared to investigate a trajectory of statistical thinking, from novice thinkers to more advanced thinkers to expert thinkers. For example, do statisticians think of more improvements to the survey in MODEST than novice students?

Future research could also gather more evidence of validity for MODEST as an assessment of statistical thinking. This could include administering MODEST to statisticians and examining their scores. If MODEST is a good assessment of statistical thinking, then experts would obtain a high overall score of statistical thinking. Additionally, a larger number of student responses could be collected to further evaluate MODEST as an assessment of statistical thinking. A factor analysis would be carried out on these results, with the goal of investigating whether MODEST measures one factor (i.e., statistical thinking).

Wild and Pfannkuch (1999) described characteristics that make up statistical thinking, which was useful for creating the test blueprint for MODEST. However, transforming Wild and Pfannkuch's descriptions into measurable outcomes was challenging. Some of their descriptions of the characteristics of statistical thinking were too vague when writing items or developing the rubric. For example, Wild and Pfannkuch described the element of *creates a model* as "constructing models and using

them to understand and predict the behaviour of aspects of the world that concern us..." (1999, p. 230). This definition was not helpful in trying to assess students' ability to create a model. As a result, other resources needed to be consulted to further elaborate on those elements of statistical thinking that were ill-defined in Wild and Pfannkuch's framework. Future work could continue to define the characteristics of statistical thinking so that there is less ambiguity in trying to measure statistical thinking. Another reason for continuing the work of Wild and Pfannkuch is to explore whether characteristics of statistical thinking have changed over the last 17 years. For example, one characteristic that may be absent from their framework is computational thinking, or thinking about programming. Given the popular new undergraduate major of Data Science in Statistics department across the U.S., it would seem that computational thinking would be a part of statistical thinking.

5.6 Conclusion

MODEST was created to try to understand students' statistical thinking in an introductory course based on modeling and simulation. Based on the data that were collected from expert reviewers and a variety of students in the pilot test phase, there is evidence suggesting that MODEST measures the construct of statistical thinking and its development in students.

Furthermore, for the introductory course used in this study, students appear to enter the course with a moderate amount of statistical thinking and leave having developed some statistical thinking as a result of the course. However, given the small change in the students' statistical thinking scores, more could be done within the course

to improve their thinking. Students should to be given more opportunities to develop expert-like thinking, such as encountering problems that force students to grapple with reasoning about statistical models and thinking that is not often emphasized in introductory statistics courses, such as curiosity.

Although there is still work to be done in assessing students' statistical thinking, MODEST can be a valuable addition to the statistics education community by filling the gap of assessing students' statistical thinking. Both statistics education researchers and instructors would benefit from using MODEST to understand statistical thinking. Assessing the important learning objective of "develop students' statistical thinking" in statistics courses will no longer be a mystery.

References

Alacaci, C. (2004). Inferential statistics: Understanding expert knowledge and its implications for statistics education. *Journal of Statistics Education,* 12(2). Retrieved from http://ww2.amstat.org/publications/jse/v12n2/alacaci.html.

American Psychological Association and the National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Society of Quality (1996). *Glossary and tables for SQC*. Milwaukee, WI: Quality Press.

American Statistical Association (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016.* The American Statistical Association (ASA). Retrieved November 23, 2016 from http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

Antony, J. (2004). Some pros and cons of Six Sigma: An academic perspective. *The TQM Magazine*, 16(4), 303-306.

Barnett, B. G. (1995). Developing reflection and expertise: Can mentors make the difference? *Journal of Educational Administration*, 33(5), 45-59.

Barrows, H. S. (2000). *Problem-based learning applied to medical education*. Springfield: Southern Illinois University Press.

Bédard, J., & Chi, M. T. (1992). Expertise. *Current Directions in Psychological Science*, 1(4), 135-139.

Ben-Zvi, D., & Friedlander, A. (1997). Statistical thinking in a technological environment. In J.B. Garfield and G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 45-55). Voorburg, The Netherlands: International Statistical Institute.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *Challenge of developing statistical literacy, reasoning, and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bereiter, C., & Scardamalia, M. (1986). Educational relevance of the study of expertise. *Interchange*, 17(2), 10-19.

Biehler, R. (1999). Discussion: Learning to think statistically and to cope with variation. *International Statistical Review*, 67(3), 259-262.

Binnie, N. (2002). Using projects to encourage statistical thinking. The 6th Proceedings of the International Conference on Teaching Statistics.

Blessing, S. & Anderson, J. R. (1996). How people learn to skip steps. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 576-598.

Bransford, J. (1994). Who ya gonna call? Thoughts about teaching problem solving. In P. Hallinger, K. Lithwood, & J. Murphy (Eds.), *Cognitive perspectives on educational leadership*. New York: Teacher's College Press.

Bransford, J., Brown, A., & Cocking, R. (2000). *How people learn: Brain, mind, experience, and school* (Expanded Edition). National Research Council. Washington, DC: National Academy Press.

Breslow, N. E. (1999). Discussion: Statistical thinking in practice. *International Statistical Review*, 67(3), 252-255.

Bursic, K. M., Shuman, L. J., & Sacre, M. B. (2011). Improving student attainment of ABET outcomes using model-eliciting activities (MEAS). In *Proceedings of the American Society for Engineering Education Annual Conference and Exposition*, Vancouver, British Columbia.

Carlson, M. P., & Bloom, I. (2005). The cyclic nature of problem solving: An emergent multidimensional problem-solving framework. *Educational Studies in Mathematics*, 58, 45-75.

Carlson, M., Larsen, S., & Lesh, R. (2003). Integrating a models and modeling perspective with existing research and practice. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics teaching, learning, and problem solving* (pp. 465-478). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Carnes, M.T., Cardella, M.E., & Diefes-Dux, H.A. (2010). Progression of student solutions over the course of a model-eliciting activity (MEA). *Proceedings of the 40th ASEE/IEEE Frontiers in Education Conference*, Washington, DC.

Chamberlin, M. (2004). Design principles for teacher investigations of student work. *Mathematics Teacher Education and Development*, 6, 52-62.

Chamberlin, M. (2005). Teachers' discussions of students' thinking: Meeting the challenge of attending to students' thinking. *Journal of Mathematics Teacher Education*, 8, 141-170.

Chan, E. C. M. (2008). Using model-eliciting activities for primary mathematics classrooms. *The Mathematics Educator*, 11(1), 47-66.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3), 1-17.

Chen, C., & Ge, X. (2006). The design of a web-based cognitive modeling system to support ill-structured problem solving. *British Journal of Educational Technology*, 37(2), 299-302.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.

Chi, M.T.H. & Glaser, R. (1985). Problem solving ability. In R.J. Sternberg (Ed.), *Human abilities: An information processing approach* (pp. 227-250). New York: W.H. Freeman.

Chick, H. L., & Watson, J. M. (2002). Collaborative influences on emergent statistical thinking – a case study. *Journal of Mathematical Behavior*, 21, 371-400.

Chin, C., & Chia, L. (2005). Problem-based learning: Using ill-structured problems in biology project work. *Science Education,* 90(1), 44-67.

Cobb, G. (1992). Teaching statistics. In Lynn A. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (MAA Notes No. 22, pp. 3-43). Washington, DC: Mathematical Association of America.

Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1). Retrieved from: http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art1

Colvin, S., & Vos, K. E. (1997). Authentic assessment models for statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 27-36). Amsterdam: IOS Press.

delMas, R.C. (2002). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *Challenge of developing statistical literacy, reasoning, and thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

DeFranco, T. C. (1996). A perspective on mathematical problem-solving expertise based on the performances of male Ph. D. mathematicians. *Research in Collegiate Mathematics Education II*, 6, 195-213.

Derry, S. J., Levin, J. R., Osana, H. P., Jones, M. S., & Peterson, M. (2000). Fostering students' statistical and scientific thinking: Lessons learned from an innovative college course. *American Educational Research Journal,* 37(3), 747-773.

Derry, S., Levin, J. R., & Schauble, L. (1995). Stimulating statistical thinking through situated simulations. *Teaching of Psychology,* 22(1), 51-57.

Diefes-Dux, H. A., Moore, T., Zawojewski, J., Imbrie, P. K., & Follman, D. (2004). A framework for posing open-ended engineering problems: Model-eliciting activities. In *CD Proceedings Frontiers in Education Conference*, F1A.3-F1A.8. Institute of Electrical and Electronic Engineers.

Diefes-Dux, H. A., Zawojewski, J. S., Hjalmarson, M. A. (2010). Using educational research in the design of evaluation tools for open-ended problems. *International Journal of Engineering Education*, 26(4), 807-819.

English, L. D., Lesh, R., Fennewald, T. (2008). Methodologies for investigating relationships between concept development and the development of problem solving abilities. In *11th International Congress on Mathematical Education*. Monterrey, Mexico.

Ertmer, P. A., Stepich, D. A., York, C. S., Stickman, A., Wu, X., Zurek, S., Goktas, Y. (2008). How instructional design experts use knowledge and experience to solve ill-structured problems. *Performance Improvement Quarterly,* 2(1), 17-42.

Fernandes, R., & Simon, H. A. (1999). A study of how individuals solve complex and ill-structured problems. *Policy Sciences*, 32, 225-245.

Gagné, R. M. (1980). *The conditions of learning*. New York: Holt, Rinehart, & Winston.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM: The International Journal on Mathematics Education*, 44(7), 883-898.

Garfield, J., & Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education* (pp. 133-145). Netherlands: Springer.

Ge, X., & Land, S. M. (2004). A conceptual framework for scaffolding ill-structured problem-solving processes using question prompts and peer interactions. *Educational Technology Research and Development*, 52(2), 5-22.

Glaser, R. & Chi, M.T. H. (1988). Overview. In M. T. H. Chi, R. Glaser & M. Farr (Eds.), *The nature of expertise* (pp. xv–xxviii). Hillsdale, NJ: Erlbaum.

Groth, R. E. (2005). An investigation of statistical thinking in two different contexts: Detecting a signal in a noisy process and determining a typical value. *Journal of Mathematical Behavior,* 24, 109-124.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Education Communication and Technology*, 29(2), 75-91.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York: Routledge.

Hardiman, P. T., Dufresne, R., Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory and Cognition,* 17(5), 627-638.

Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics*, 60(7), 627-636.

Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education,* 28(5), 524-549.

Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 215-238). Hillsdale, NJ: Erlbaum.

Hjalmarson, M. A., Moore, T. J., & delMas, R. (2011). Statistical analysis when the data is an image: Eliciting student thinking about sampling and variability. *Statistics Education Research Journal*, 10(1), 15-34.

Ho, C. (2000). Some phenomena of problem decomposition strategy for design thinking: differences between novices and experts. *Design Studies*, 22(1), 27-45.

Hoerl, R.W. (2001). Six sigma black belts: what do they need to know? *Journal of Quality Technology*, 33(4), 391-406.

Hoerl, R. W., & Snee, R. D. (2002). *Statistical thinking: Improving business performance.* Pacific Grove, CA: Duxbury Press.

Hoerl, R. W., & Snee, R. D. (2010). Statistical thinking and methods in quality improvement: A look to the future. *Quality Engineering*, 22(3), 119–129.

Hogan, T.P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, 20(4), 427-441.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, 45(1), 65-94.

Jonassen, D. H., & Hernandez-Serrano, J. (2002). Case-based reasoning and instructional design: Using stories to support problem solving. *Educational Technology Research and Development*, 50(2), 65-77.

Jones, G. A, Langrall, C. W., Thornton, C. A., Mooney, E. S., Wares, A., Jones, M. R., Perry, B., Putt, I. J., & Nesbet, S. (2001). Using students' statistical thinking to inform instruction. *Journal of Mathematical Behavior*, 20, 109-144.

King, F.J., Goodson, L., and Rohani, F. (1998) Higher-Order Thinking Skills: Definitions, Teaching Strategies, and Assessment. Retrieved from http://www.cala.fsu.edu/files/higher_order_thinking_skills.pdf.

Kitchener, K. S. (1983). Cognition, metacognition, and epistemic cognition: A three-level model of cognitive processing. *Human Development*, 26(4) , 222-232.

Korf, R. E. (1985). Macro-operators: A weak method for learning. *Artificial Intelligence*, 26, 35-77.

Kramer, D. A., & Woodruff, D. S. (1986). Relativistic and dialectical thought in three adult age-groups. *Human Development*, 29(5), 280-290.

Kuhn, D. (1991). *The skills of argument*. New York: Cambridge University Press.

Larkin, J. H. D., McDermott, D., Simon, D.P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science*, 4, 317-345.

Lesh, R., Amit, M., Schorr, R. Y. (1997). Using "real-life" problems to prompt students to construct conceptual models for statistical reasoning. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 65-83). Amsterdam: IOS Press.

Lesh, R., Cramer, K., Doerr, H. M., Post, T., & Zawojewski, J. S. (2003). Model development sequences. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 35- 58). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Doerr, H. M. (2003). Foundations of a models and modeling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics teaching, learning, and problem solving* (pp. 3-33). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lesh, R., & Harel, G. (2003). Problem solving, modeling, and local conceptual development. *Mathematical Thinking and Learning*, 5(2&3), 157-189.

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought-revealing activities for students and teachers. In A. Kelly & R. Lesh (Eds*.), Handbook of research in mathematics and science education* (pp. 113-149). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Kelly, A. E. (1994). Action-theoretic and phenomenological approaches to research in mathematics education: Studies on continually developing experts. In R. Biehler, R. W. Scholz, R. Sträβer, & B. Winkelmann (Eds.), *Didactics of mathematics as a scientific discipline* (pp. 277-286). Dordrecht: Kluwer Academic Publishers.

Lesh, R., & Kaput, J. (1988). Interpreting modeling as local conceptual development. In J. DeLange & M. Doorman (Eds.), *Senior secondary mathematics education.* Utrecht, Netherlands: OW&OC.

Lesh, R., & Zawojewski, J. (2007). Problem solving and modeling. *Second handbook of research on mathematics teaching and learning*, 2, 763-804.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the power of data.* Hoboken, NJ: Wiley.

MacGillivray, H., & Pereira-Mendoza, L. (2011). Teaching statistical thinking through investigative projects. In C. Batanero, G. Burrill, & C. Reading (Eds), *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education* (pp. 109-120). The Netherlands: Springer.

MacKay, B. (n.d.) http://serc.carleton.edu/introgeo/conceptmodels/why.html.

MacKay, J.M., & Elam, J.J. (1992). A comparative study of how experts and novices use a decision aid to solve problems in complex knowledge domains. *Information systems Research*, 3(2), 150-172.

Mackay, R.J. & Oldford, W. (1994). *Stat 231 Course Notes Fall 1994*. Waterloo: University of Waterloo.

Makar, K., & Confrey, J. (2002). Comparing two distributions: Investigating secondary teachers' statistical thinking. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*. Cape Town, South Africa: International Association for Statistics Education.

Measuring Study Effectiveness (2013). In *Pedagogy in Action*: *the SERC Portal for Educators*. Retrieved June 10, 2014 from https://serc.carleton.edu/sp/library/mea/examples/example1.html.

Melton, K. I. (2004). Statistical thinking activities: Some simple exercises with powerful lessons. *Journal of Statistics Education*, 12(2), 1-10.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Misfeldt, M., & Johansen, M.W. (2015). Research mathematicians' practices in selecting mathematical problems. *Educational Studies in Mathematics*, 89(3), 357-373.

Moore, D. S. (1990). Uncertainty. In L.A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-138). Washington, D. C.: National Academy Press.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, 65(2), 123-165.

Moore, D. (1999). Discussion: What shall we teach beginners?. *International Statistical Review*, 67(3), 250-252.

Moore, T. J., & Hjalmarson, M. A. (2010). Developing measures of roughness: Problem solving as a method to document student thinking in engineering. *International Journal of Engineering Education*, 26(4), 820-830.

Moore, T. J., Miller, R. L., Lesh, R. A., Stohlmann, M. S., & Kim, Y. R. (2013). Modeling in engineering: The role of representational fluency in students' conceptual understanding. *Journal of Engineering Education*, 102(1), 141-178.

National Governors Association Center for Best Practices and the Council of Chief State School Officers (2010). *Common Core State Standards for Mathematics*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Park, J. (2012). Developing and validating an instrument to measure college students' inferential reasoning in statistics: An argument-based approach to validation. (Doctoral Dissertation, University of Minnesota, 2012). Retrieved from http://iase-web.org/Publications.php?p=Dissertations.

Pearson. (2011). *Metacognition: A literature review; Research report.* Upper Saddle River: NJ: E. Lai.

Perkins, D. N., Faraday, M. and Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal reasoning and education* (pp. 83-106). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pfannkuch, M., & Horring, J. (2005). Developing statistical thinking in a secondary school: A collaborative curriculum development. In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education: International Association for Statistical Education (IASE) Roundtable*, Lund, Sweden, 28 June-3 July 2004, (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.

Pfannkuch, M., & Rubick, A. (2002). An exploration of students' statistical thinking with given data. *Statistics Education Research Journal*, 1(2), 4-21.

Pfannkuch, M., & Wild, C. J. (2000). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statistical Science*, 15(2), 132-152.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *Challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pfannkuch, M., & Wild, C. J. (2003). Statistical thinking: How can we develop it? In Bulletin of the International Statistical Institute 54th Session Proceedings, Berlin, 2003 [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Reitman, W. (1965). *Cognition and thought*. New York: Wiley.

Resnick, L. (1989). Treating mathematics as an ill-structured discipline. In R. Charles & E. Silver (Eds.), The teaching and assessing of mathematical problem solving, pp. 32-60. Reston, VA: National Council of teachers of Mathematics.

Sanchez, E., & Blancarte, A.L.G. (2008). Training in-service teachers to develop statistical thinking. The 8th Proceedings of the International Conference on Teaching Statistics.

Schenk, K.D., Vitalari, N.P., & Davis, K.S. (1998). Differences between novice and expert systems analysts: What do we know and what do we do? *Journal of Management Information Systems*, 15(1), 9-50.

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading, 3*, 257–280.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), Handbook for research on mathematics teaching and learning (pp. 334-370). New York: MacMillan.

Schoenfeld, A. H. (1998). Making mathematics and making pasta: From cookbook procedures to really cooking. In J. G. Greeno & S. Golman (Eds.), *Thinking practices: A symposium on mathematics and science learning* (pp. 299-319). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 8(5), 484-494.

Schraw, G., Dunkle, M. E., & Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem-solving. *Applied Cognitive Psychology,* 9, 523-538.

Sedlmeier, P. (2000). How to improve statistical thinking: Choose the task representation wisely and learn by doing. *Instructional Science*, 28, 227-262.

Shin, N., Jonassen, D. H., McGee, S. (2003). Predictors of well-structured and ill-structured problem solving in an astronomy simulation. *Journal of Research in Science Teaching,* 40(1), 6-33.

Sinnott, J. D. (1989). A model for solution of ill-structured problems: Implications for everyday and abstract problem solving. In J. D. Sinnott (Ed.), *Everyday problem solving: Theory and applications* (pp. 72-99). New York: Praeger.

Smith, T. M. F. (1999). [Statistical thinking in empirical enquiry]: Discussion. *International Statistical Review*, 67(3), 248-250.

Snee, R. D. (1990). Statistical thinking and Its contribution to total quality. *The American Statistician*, 44(2), 116-121.

Snee, R. D. (1993). What's missing in statistical education? *The American Statistician*, 47(2), 149-154.

Snee, R. D. (1999). Discussion: Development and use of statistical thinking: A new era. *International Statistical Review*, 67(3), 255-258.

Spector, J. M. (2006). A methodology for assessing learning in complex and ill-structured task domains. *Innovations in Education and Teaching International*, 43(2), 109-120.

Sternberg, R. J. (1996). What is mathematical thinking? In R. J. Sternberg & T. Ben-Zeev (Eds), *The nature of mathematical thinking* (pp. 303-318). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stylianou, D. A. (2002). On the interaction of visualization and analysis: the negotiation of a visual representation in expert problem solving. *Journal of Mathematical Behavior,* 21, 303-317.

Stylianou, D. A., & Silver, E. A. (2004). The role of visual representations in advanced mathematical problem solving: an examination of expert-novice similarities and differences. *Mathematical Thinking and Learning,* 6(4), 353-387.

Tintle, N. Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2013). Introduction to statistical investigations. Retrieved November 23, 2016 from http://www.math.hope.edu/isi/

van Someren, M.W., Barnard, Y.F., & Sandberg, J.A.C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London, England: Academic Press.

Voss, J. F. (2006). Toulmin's model and the solving of ill-structured problems. In D. Hitchcock & B. Verheij (Eds.), *Arguing on the Toulmin Model: New essays in argument analysis and evaluation* (pp. 303-311). Springer.

Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum.

Voss, J. F., Tyler, S. W., & Yengo, L. A. (1983). Individual differences in the solving of social science problems. In R. F. Dillon & R. R. Schmeck (Eds.), *Individual differences in cognition* (pp. 204-232). New York: Academic Press.

Voss, J. F., Wolfe, C. R., Lawrence, J. A., & Engle, R. A. (1991). From representation to decision: An analysis of problem solving in international relations. In R. J. Sternberg, & P. A. Frensch (Eds.), *Complex problem solving* (pp. 119-158). Hillsdale, NJ: Lawrence Erlbaum.

Watson, J.M. (1997). Assessing statistical thinking using the media. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge of statistics education* (pp. 107-121). Amsterdam, Netherlands: IOS Press.

Weiss, R. E. (2003). Designing problems to promote higher-order thinking. *New Directions for Teaching and Learning,* 95, 25-31.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223-248.

Wild, C.J. (1994). Embracing the "wider view" of statistics. *American Statistician,* 48(2), 163-171

Woods, D. R., Hrymak, A. N., Marshall, R. R., Wood, P. E., Crowe, C. M., Hoffman, T. W. Wright, J. D., Taylor, P. A., Woodhouse, K. A., & Bouchard, C. G. K. (1997). Developing problem solving skills: The McMaster Problem Solving program. *Journal of Engineering Education*, 86(2), 75-91.

Yildirim, T. P., Shuman, L., & Besterfield-Sacre, M. (2010). Model-eliciting activities: Assessing engineering student problem solving and skill integration processes. *International Journal of Engineering Education*, 26(4), 831-845.

Ziegler, L. (2014). Reconceptualizing statistical literacy: Developing an assessment for the modern introductory statistics course (Doctoral dissertation, University of Minnesota). Retrieved from http://iase-web.org/documents/dissertations/14.LauraZiegler.Dissertation.pdf.

Appendix A
Versions of MODEST

Appendix A1: Original Measuring Study Effectiveness MEA

Limerick Post on the Web
Effective study skills for getting best exam results
Thursday, 06 August 2009 09:09

As students prepare to return to school, Julie Kilmartin, principal of Kilmartin Educational Services, stresses the importance of getting off to a good September start by ensuring that students have the necessary skills required to maximize study time. Therefore, Kilmartin Educational Services offers students preparation for the return to school. Indeed, demand for the August Preparation Course students is particularly high this year.

On Saturday August 29, students may participate in a study skills seminar which is especially designed to help students organize their study and homework as they return to school.

According to Louise Brett, course coordinator, the study seminar will provide students with an insight into how they should approach and plan their school work over the coming months. Emphasis will be placed on a number of key topics:

* Maximizing Class Time-Taking and Making Notes

Emphasis will be placed on the importance of creating effective class and how to organize notes. Summary notes play a vital role during exam revision.

* Devising Realistic Study Plans

It is important that students plan their homework and revision sessions. Revision must be effective and organized and during this seminar, students will learn how to create effective, realistic study plans.

* Dealing with Homework

As always, homework is an integral part of any successful exam result. Students will be advised as how to handle homework as part of an effective study schedule.

* Memory Techniques

In our first seminar, students will be given a number of techniques which will help students retain information and recall it with ease during exams.

* Exam Techniques

A brief reference shall be made to exams and some of the various exam techniques which will be invaluable to students.

* Study and Staying Healthy

Exams are important but staying healthy during the preparation for exams is equally important.

All in all, this seminar on Saturday August 29 will get students to focus on successful methods that help students organize their studies and improve their grades in school.

Readiness Questions

1. How effective are your study habits?

2. How would you describe a student who has effective study habits?

3. What do you think are the most important study habits needed by students to succeed in school?

4. What types of questions should be used to determine how effective a student studies?

Once everyone in your group has answered all 4 questions, share and discuss your responses to each question.

Problem Statement

Background: Measuring Student's Effective Study Habits

Students in an education course developed a questionnaire to help themselves and future students self-assess their own study habits for the course, The Survey of Study Effectiveness. Their hope was that future students could take the questionnaire about halfway through the semester; and determine how effectively they are studying and perhaps make improvements to their study habits if needed. The survey that the students developed is presented on the next page. Familiarize yourself with the survey by reading through it, and then go on to the group task presented below.

Developing an Index of Study Effectiveness

Questionnaires given to people so they can rate themselves on some characteristic usually offer some kind of summary graph or score that provides a direct answer to questions such as "how good a friend are you" or "how smart are you", etc. For the questionnaire about study effectiveness, a summary score can be thought of as an index of study effectiveness.

Please write a report to the students in the Education Course explaining how results from the questionnaire they developed can be used to produce the index of study effectiveness that can be used to judge how well students are studying. Please include any formulas, procedures, or rules on which your index of study effectiveness is based. As part of your report, use the information in the table below to explain and illustrate how your method works for these five students. What score for "study effectiveness" does your method assign to each? Use your index to put these students in rank order according to their study effectiveness.

As you develop your index of study effectiveness, remember that the questions on the questionnaire were intended to provide information about the issues named in the column labeled Topic in the table.

A Sample of Five Student's Responses to the Questionnaire

	Topic	Al	Barbara	Carl	Deborah	Ed
Question 1	**Difficulty of Assignments**	2	2	4	4	3
Question 2	**Prior Knowledge**	4	2	1	5	4
Question 3	**Scores on Quizzes**	4	4	4	5	2
Question 4	**Achievements not on Tests**	1	2	4	3	2
Question 5	**Time Spent Studying**	3	1	4	5	1
Question 6	**Quality of Time Studying**	3	2	1	3	4
Question 7	**Skipping Parts**	2	3	4	3	2
Question 8	**Re Reading & Reorganizing**	2	3	1	2	2
Question 9	**Skim for Big Picture**	4	4	1	3	4

Question 10	**Underline, Outline Review**	1	2	3	5	2
Question 11	**Reorganize & Reword**	3	4	4	4	2
Question 12	**Finding Personal Examples**	2	3	3	4	2
Question 13	**Explain & Simplify**	2	1	2	5	3
Question 14	**Identify Big Ideas**	3	0	2	5	3

Students Study Effectiveness Survey

1. Compared to other similar courses that you have taken, what is the difficulty level of the topics you are studying in this course?

   Much Lower | Lower | About the Same | Higher | Much Higher

2. Compared to other students in the class, did you already know a lot about the topics we are studying?

   A lot less | A little less | About the same | A little more | A lot more

3. Compared to other students in the class, how are you doing on quizzes over homework assignments/

   much worse | a little worse | About the same | A little better | A lot better

4. Do you think that your scores on quizzes and tests are accurate assessments of how well you have learned topics you've studied in this course.

   I'm not doing as well as the quizzes suggest | I'm not doing a little less well than the quizzes suggest | I'm doing about as well as the quiz scores suggest | I'm doing a little better than the quiz scores suggest | I'm doing a lot better than the quiz scores suggest

5. How much time do you usually spent studying to prepare for each class?

   Less than 30 minutes | Less than an hour | Between one hour and two | More than 2 hours | More than three hours

6. How much are you distracted when you are studying for each class?

   Never | Once or Twice | Sometimes | Regularly | Very Often

7. Do you usually skip parts of assignments or things that are important to do?

   Never | Sometimes | I do exactly what was assigned. | I go beyond what was assigned. | I go a lot beyond what was assigned.

8. Do you only ready things once, or do you read several times?

   I never read things more than once | I sometimes read things more than once | About half of the read things more than once. | I usually read things more than once | I always read things more than once.

9. Before reading the course material, how often do you survey the chapter to develop a general idea of what the reading will be about?

   Never | Once or Twice | Sometimes | Regularly | Very Often

10. While reading, how often do you take notes, highlight, mark in the margins, ask and answer questions about the material?

Never	Once or Twice	Sometimes	Regularly	Very Often

11. When reading the course material, how often do you sacrifice comprehension for just getting through the pages assigned?

Never	Once or Twice	Sometimes	Regularly	Very Often

12. Do you reorganize what you have read - putting it into your own words, and using your own examples to emphasize the main points?

Never	Once or Twice	Sometimes	Regularly	Very Often

13. How often do you discuss things with others – trying to simplify, summarize, reorganize, and things you are studying.

Never	Not often	Usually some	Usually quite a lot	Always and often

14. After you read about some new topic, how often do you try to relate the main "big ideas" to those emphasized in other topics you have studied in the course?

Never	Not often	Usually some	Usually quite a lot	Always and often

Appendix A2: MODEST (version 1)

Modeling To Elicit Statistical Thinking
(MODEST 1)[1]

INSTRUCTIONS

The purpose of this assessment is to evaluate how you think about a problem from a statistical point of view.

Directions:

2. Read the brief article to help you familiarize yourself with the problem scenario that you will be investigating.

3. Answer the questions related to solving the problem.
   - The questions are open-ended questions, so provide as much detail in your answers as possible so someone else can follow your thinking.
   - You will be evaluated based on how you describe your thought processes in your answers.
   - You may want to have a piece of paper and writing utensil available while you are solving the problem.

[1] Measuring Study Effectiveness (2013). In Pedagogy in Action: the SERC Portal for Educators. Retrieved July 15, 2014 from http://serc.carleton.edu/sp/library/mea/examples/example1.html.

ONLINE NEWS ARTICLE

Effective study skills for getting best exam results[2]

As students prepare to return to school, it is important to get off to a good September start by ensuring that students have the necessary skills required to maximize study time. Students may participate in a study skills seminar that is especially designed to help students organize their study and homework as they return to school. Study skills seminars typically focus on a number of key topics:

Maximizing Class Time-Taking and Making Notes

Students learn the importance of creating an effective class and learn how to organize notes. Summary notes play a vital role during exam revision.

Devising Realistic Study Plans

Students learn how to create effective, realistic study plans for their homework and revision sessions.

Dealing with Homework

Since homework is an integral part of any successful exam result, students are advised on how to handle homework as part of an effective study schedule.

Memory Techniques

Students are given multiple memory techniques to help in the retention and recollection of information during exams.

Exam Techniques

Students are introduced to invaluable exam-taking techniques.

Study and Staying Healthy

Students learn the importance of studying and staying healthy while preparing for exams.

All in all, study skills seminars get students to focus on successful methods that help them organize their studies and improve their grades in school.

[2] *Note.* Adapted from Effective study skills for getting best exam results (2009, Aug 6). Retrieved on July 3, 2014 from http://www.limerickpost.ie/2009/08/06/effective-study-skills-for-getting-best-exam-results/.

Task

Suppose you work for Westat as a survey design and analysis expert. Tutors for a study skills seminar hired you to create a survey to help students self-assess their own study habits for a course. Their hope was that students could take the survey about halfway through the semester and determine how effectively they are studying and perhaps make improvements to their study habits if needed.

PART I: Plan

Imagine that you were about to create the survey for the tutors.

1. What student or class characteristics would be useful on the survey to determine the effectiveness of students' study habits? For each characteristic, explain why you would include it on the survey.

2. Do you think all of the characteristics you listed are equal with respect to helping a student be an effective studier? Explain your reasoning.

3. Would you ask the tutors questions about their knowledge of students' study habits to help you create the survey? Why or why not?

Let's say you created the survey and called it the *Study Effectiveness Survey*. See the following link to familiarize yourself with the questions and format of the *Study Effectiveness Survey*: [*Note. See pages 7-8]*

PART II: Develop and Test a Measure of Study Effectiveness

Your first task is to develop a summary score of study effectiveness that will be used on the students' responses to the survey. This score will help the students judge how effective their study habits are. Please refer to the survey and the information in the table below to help you come up with your score.

A Sample of Five Student's Responses to the Study Effectiveness Survey

	Question Content	Al	Barbara	Carl	Deborah	Ed
Question 1	Difficulty of Topics	Lower	Lower	Higher	Higher	About the same
Question 2	Prior Knowledge	A little more	A little less	Much less	Much more	A little more

Question 3	Scores on Assignments	A little better	A little better	A little better	Much better	A little worse
Question 4	Grades	Not at all	Very little	To some extent	Fair	Very little
Question 5	Time Spent Studying	60-120 minutes	0-30 minutes	120-180 minutes	180 minutes or more	0-30 minutes
Question 6	Quality of Time Studying	Sometimes	Very rarely	Never	Sometimes	Regularly
Question 7	Skipping Parts	Very rarely	Sometimes	Regularly	Sometimes	Very rarely
Question 8	Amount of Reading	Very rarely	Sometimes	Never	Very rarely	Very rarely
Question 9	Skim for Big Picture	Regularly	Regularly	Never	Sometimes	Regularly
Question 10	Notetaking when Reading	Never	Very rarely	Sometimes	Very often	Very rarely
Question 11	Quality of Reading	Sometimes	Regularly	Regularly	Regularly	Very rarely
Question 12	Synthesize the Readings	Very rarely	Sometimes	Sometimes	Regularly	Very rarely
Question 13	Discuss with Others	Very rarely	Never	Very rarely	Very often	Sometimes
Question 14	Make Connections	Sometimes	Never	Very rarely	Very often	Sometimes

4. Report how to produce your score of study effectiveness given a student's result to the survey. Please include any formulas, procedures, or rules on which your score of study effectiveness is based.

5. Explain your thought process that occurred while coming up with your score (e.g., what did you consider, what did you notice, what previous knowledge did you use).

6. For each of the five students in the table, calculate and report their score of study effectiveness using your method described earlier. Also, place these students in rank order according to their study effectiveness.

7. If two students have the same score of study effectiveness, does that mean they have the same study habits, according to the content on the survey? Explain your reasoning.

PART III: Evaluation

8. Write a report to the tutors of the study skills seminar that describes

- how to calculate your score of study effectiveness using the *Study Effectiveness Survey* and
- how to interpret the score so students can judge how effective their study habits are.

9. Do you think it would be useful to also provide students with a summary graph of their results on the survey? If so, what type of graph would you present and how would this be useful to the students? If no, why not?

10. Having done this activity, do you have concerns or reservations about using the *Study Effectiveness Survey* or your score of study effectiveness within a classroom? Explain.

11. Are there other student or classroom characteristics that you would want to examine to help explain effective study habits? Explain your reasoning.

12. Did you wonder what student population the *Study Effectiveness Survey* would apply to? If so, what were your thoughts? If not, why not?

Study Effectiveness Survey

Difficulty of Topics:

1. How does the difficulty level of the topics in this course compare to other similar courses you have taken?

Much lower	Lower	About the same	Higher	Much higher

Prior Knowledge:

2. How does your prior knowledge about the topics in this course compare to other students in the class?

Much less	A little less	About the same	A little more	Much more

Scores on Assignments:

3. How do your assignment scores compare to other students in the class?

Much worse	A little worse	About the same	A little better	Much better

Grades:

4. How well do you think your current grade indicates how much you have learned in this course?

Not at all	Very little	Fair	To some extent	To a great extent

Time Spent Studying:

5. Approximately how much time each week do you spend studying to prepare for this class?

0-30 minutes	30-60 minutes	60-120 minutes	120-180 minutes	180 minutes or more

Quality of Time Studying:

6. How often are you distracted when you are studying for this class?

Never	Very rarely	Sometimes	Regularly	Very often

Skipping Parts:

7. How often do you skip parts of assignments or things that are important to do?

Never	Very rarely	Sometimes	Regularly	Very often

Amount of Reading:

8. How often do you read the course material more than once?

Never	Very rarely	Sometimes	Regularly	Very often

Skim for Big Picture:

9. Before reading the course material, how often do you survey the chapter to develop a general idea of what the reading will be about?

Never	Very rarely	Sometimes	Regularly	Very often

Notetaking when Reading:

10. When reading the course material, how often do you take notes, highlight text, mark in the margins, or ask questions about the material?

Never	Very rarely	Sometimes	Regularly	Very often

Quality of Reading:

11. When reading the course material, how often do you sacrifice comprehension for just getting through the pages assigned?

Never Very rarely Sometimes Regularly Very often

Synthesize the Readings:

12. When reading the course material, how often do you put what you've read into your own words or used your own examples to emphasize the main points?

Never Very rarely Sometimes Regularly Very often

Discuss with Others:

13. How often do you discuss the topics with other students in this course (e.g., try to simplify, summarize, or reorganize topics)?

Never Very rarely Sometimes Regularly Very often

Make Connections:

14. How often do you try to make connections between the topics learned in the course?

Never Very rarely Sometimes Regularly Very often

Appendix A3: MODEST (version 4)

Modeling To Elicit Statistical Thinking

(MODEST 4)[3]

INSTRUCTIONS

The purpose of this assessment is to evaluate how you think about a problem from a statistical point of view.

Directions:

1. Read the brief article to help you familiarize yourself with the problem scenario that you will be investigating.

2. Answer the questions related to solving the problem.
   - The questions are open-ended questions. Provide as much detail in your answers as possible so someone else can follow your thinking.
   - You will be evaluated based on how you describe your thought processes in your answers.
   - You may want to have writing materials available while you are solving the problem.

[3] Measuring Study Effectiveness (2013). In Pedagogy in Action: the SERC Portal for Educators. Retrieved July 15, 2014 from http://serc.carleton.edu/sp/library/mea/examples/example1.html.

Directions: Read the brief online news article to help you familiarize yourself with the problem scenario.

ONLINE NEWS ARTICLE

Effective Study Skills[4]

This web page on study skills, and what some may refer to as study tips, is designed to help you improve your learning and understanding, and ultimately your grades.

Here are some general techniques that seem to produce good results.

Strategies

When to study. The problem of when to study is critical. A good rule of thumb is that studying should be carried out only when you are rested, alert, and have planned for it.

Studying for lecture courses. If your study period is before the lecture class, be sure you have read all the assignments and made notes on what you don't understand. If the study period is after the lecture class, review the notes you took during class while the information is still fresh.

Thinking skills. If you're not a good thinker, start now by developing habits that make you ask yourself questions as you read. Talk to other students who you feel are good thinkers. Ask them what it is they do when they think critically or creatively.

The SQ3R method. SQ3R stands for Survey, Question, Read, Recite, Review.

Survey - get the best overall picture of what you're going to study BEFORE you study it in any detail.

Question - ask questions for learning.

Read - When you read, read actively. Read to answer questions you have asked yourself or questions the instructor or author has asked.

Recite - When you recite, you stop reading periodically to recall what you have read.

Review - A review is a survey of what you have covered. It is a review of what you are supposed to accomplish, not what you are going to do.

Taking Lecture Notes

Surveying, Questioning, Listening. Your main job in taking lecture notes is to be a good listener. To be a good listener, you must learn to focus and concentrate on the main points of the lecture. Get them down, and then later reorganize them in your own words.

Reviewing and Revising. Begin the process by reviewing (and potentially revising) your notes right after a lecture. If you wait too long, you may discover that the notes just don't make sense.

A Final Word

The study skills presented here depend on one thing, and that is your willingness to WANT to improve and do well in school. You are the one who is responsible for your education, and effective study skills can help you. To that end, one last word of advice -- work smart, not hard.

[4] *Note.* Adapted from Kilzik, B. (2014, Sept. 30). Effective study skills. Retrieved on October 15, 2014 from http://www.adprima.com/studyout.htm.

Directions: Answer the questions related to solving the problem.

- The questions are open-ended questions. Provide as much detail in your answers as possible so someone else can follow your thinking.
- You will be evaluated based on how you describe your thought processes in your answers.
- You may want to have writing materials available while you are solving the problem.

Task

Suppose you work as a survey design and analysis expert. An instructor of a college math course has ~~Tutors for a study skills seminar have~~ hired you to create a survey that instructors can use to evaluate their students' study habits ~~helps students self-assess their own study habits for a course~~. The goal is to have students take the survey about halfway through the semester. The results of the survey would be used to determine how effectively students are studying and to make improvements to their study habits, if needed.

Changes: Based on the feedback from the expert reviewers to better specific the client, the section description was modified to better clarify the problem.

PART I: Plan

Imagine that you were about to create the student study habits survey ~~survey for the tutors~~.

Changes: The section description was modified to better clarity the problem.

1. ~~List student or class characteristics that you think would be useful for determining the effectiveness of students' study habits.~~ Use the article provided, your own experience, and/or other external resources ~~for creating your list.~~ to create a list of factors that might affect student study habits (such as student characteristics, course characteristics, environment characteristics, etc.). For each characteristic, briefly explain why it would be useful when examining student study habits ~~you would include it on the survey~~.

Changes: Based on the feedback from the expert reviewers, who found this item confusing, the item was revised to clarify what is being asked in the item.

2. (NEW ITEM to assess the element of *considers variation*) Do you think that the effect of the study habit factors on course grade would be the same for all students in the course? Explain why the factors would or would not have the same effect on course grade for all students.

(OLD ITEM from MODEST 3 to assess the element of *considers variation*) ~~Are the characteristics you listed of equal value with respect to helping a student be an effective studier? Explain your reasoning.~~

Changes: Based on the low ratings from the expert reviewers and the feedback that the item does not capture the essence of *considers variation*, the item was rewritten to better measure the statistical thinking element of *considers variation*.

3. (NEW ITEM to assess the elements of *develops a plan* and *is open to new ideas*) In order to create the survey and use it to evaluate student study habits, what other kinds of information would you need to consider? Be sure to explain your reasoning.

(OLD ITEM from MODEST 3 to assess the elements of *develops a plan* and *is open to new ideas*) ~~Would asking the tutors questions about their knowledge of students' study habits be useful when creating the survey? Why or why not?~~

Changes: Based on the feedback from the expert reviewers that the item was limiting in eliciting students innate thought processes, the item was rewritten to be a more open-ended, ill-structured question.

Let's say ~~you created the survey and called it the~~ the instructor obtained a preexisting survey, called *Study Effectiveness Survey,* from a colleague and collected data from five of her students. See ~~the following link~~ the handouts provided to familiarize yourself with the questions and format of the *Study Effectiveness Survey*.

Changes: The section description was revised to clarify the problem and task.

PART II: Develop and Use ~~Test~~ a Measure of Study Effectiveness

Your first task is to develop ~~a summary score~~ an overall score of study effectiveness that will be used on the students' responses to the survey. This score will help ~~the students~~ instructors judge how effective their students' study habits are. *Be sure to refer to the survey and the information in the table below to help you come up with your score.*

Changes: The section description was revised to clarify the problem and task.

A Sample of Five Student's Responses to the Study Effectiveness Survey

Question	Question Content	Al	Barbara	Carl	Deborah	Ed
1	Difficulty of Topics	Lower	Lower	Higher	Higher	About the same
2	Prior Knowledge	A little	A little less	Much less	Much	A little

		more			more	more
3	Scores on Assignments	A little better	A little better	A little better	Much better	A little worse
4	Grades	Not at all	Very little	To some extent	Fair	Very little
5	Time Spent Studying	60-120 minutes	0-30 minutes	120-180 minutes	180 minutes or more	0-30 minutes
6	Quality of Time Studying	Sometimes	Very rarely	Never	Sometimes	Regularly
7	Skipping Parts	Very rarely	Sometimes	Regularly	Sometimes	Very rarely
8	Read More than Once	Very rarely	Sometimes	Never	Very rarely	Very rarely
9	Skim for Big Picture	Regularly	Regularly	Never	Sometimes	Regularly
10	Notetaking when Reading	Never	Very rarely	Sometimes	Very often	Very rarely
11	Read and Not Comprehend	Sometimes	Regularly	Regularly	Regularly	Very rarely
12	Synthesize the Readings	Very rarely	Sometimes	Sometimes	Regularly	Very rarely
13	Discuss with Others	Very rarely	Never	Very rarely	Very often	Sometimes
14	Make Connections	Sometimes	Never	Very rarely	Very often	Sometimes

4. To aid in the process of developing ~~a summary score~~ an overall score of study effectiveness, fill in the table below describing how ~~each question will contribute to the~~ you will use each question in creating an overall score ~~of study effectiveness~~. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the score.

Changes: The item was revised to clarify what is being asked in the item.

	Question Content	
Question 1	Difficulty of Topics	
Question 2	Prior Knowledge	
Question 3	Scores on Assignments	
Question 4	Grades	
Question 5	Time Spent Studying	
Question 6	Quality of Time Studying	
Question 7	Skipping Parts	
Question 8	Read More than Once	
Question 9	Skim for Big Picture	
Question 10	Notetaking when Reading	

Question 11	Read and Not Comprehend	
Question 12	Synthesize the Readings	
Question 13	Discuss with Others	
Question 14	Make Connections	

5. Report how to produce your ~~summary~~ overall score of study effectiveness given any student's responses to the survey. Be sure to include any formulas, procedures, or rules on which your score of study effectiveness is based.

Changes: The item was revised to clarify what is being asked in the item.

6. ~~For each of the five students in the table~~ Using the instructor's data for her five students, calculate and report ~~their~~ each student's overall score of study effectiveness using your method described earlier. ~~Also, place these students in rank order according to their study effectiveness.~~ A table is provided below to help you with this process.

Changes: The item was revised to clarify what is being asked in the item and, based on the feedback from expert reviewers about an unnecessary task in the item, the ranking aspect was removed from the item.

	Al	Barbara	Carl	Deborah	Ed
Question 1					
Question 2					
Question 3					
Question 4					
Question 5					
Question 6					
Question 7					
Question 8					
Question 9					
Question 10					
Question 11					
Question 12					
Question 13					
Question 14					
Score					

7. If two students have the same score of study effectiveness, does that mean they have the same study habits, according to the content on the survey? Explain your reasoning.

8. (NEW ITEM to assess the element of *tranforms the raw data into an aggregate form*) If the instructor collected data from her whole class, describe the statistical measures or methods that you would use to summarize how effective the instructor's students study habits are.

(OLD ITEM 9 from MODEST 3 that assessed a similar element as new item 8) ~~Do you think it would be useful to provide students with a summary graph of their results on the survey? If so, what type of graph would you present and how would this be useful to the students? If not, why not?~~

Changes: Based on the feedback from the expert reviewers about having students consider graphical aspects to analyze the data, the item was created to better measure the element of *transforms the raw data into an aggregate form*.

PART III: Evaluation

Now that you have developed and used a measure of study effectiveness, your next task will be communicating the results to the instructor and evaluating and critiquing the process of measuring study effectiveness.

Changes: The section description was added to prepare respondents on the task in Part III of the assessment.

9. (NEW ITEM to assess the elements of *draws a conclusion, integrates the statistical and contextual information, reasons with statistical models, and is skeptical*) Write a brief report (~1-2 paragraphs) to the ~~tutors of the study skills seminar~~ instructor that ~~describes~~ addresses:

    - how ~~to calculate your summary score~~ the overall score of study effectiveness was calculated using the *Study Effectiveness Survey,* and
    - ~~how to interpret the summary score so individual students can judge how effective their study habits are.~~
    - a summary of the how effective the study habits of the instructor's five students are.
    - the potential limitations of the *Study Effectiveness Survey*.
    - the potential limitations of the overall score of study effectiveness.

(OLD ITEM from MODEST 3 to assess the element of *is skeptical*) ~~What concerns or reservations would you have about using the *Study Effectiveness Survey* or your score of study effectiveness within a classroom? Explain.~~

Changes: Based on the feedback from the expert reviewers about writing concluding remarks, the item was revised to better incorporate summarization of the data and reflecting on the limitations of the tools.

10. (NEW ITEM to assess the element of *seeks alternative explanations*) What suggestions do you have to help the instructor revise the *Study Effectiveness Survey?*

(OLD ITEM from MODEST 3 to assess the element of *seeks alternative explanations*) ~~Are there other student or classroom characteristics you would consider examining to help refine your survey of effective study habits? Explain your reasoning.~~

Changes: Based on the feedback from the expert reviewers about trying to elicit a statistical thinking element without directly telling respondents to and about clarifying the phrasing in the question, the item was revised to better elicit the element of *seeks alternative explanations*.

11. (NEW ITEM to assess the element of *is curious and aware*) As a result of reading the article and working through the first 10 questions, was there anything that you wondered about regarding the evaluation of student study habits? If so, what did you wonder about? If not, why not?

(OLD ITEM from MODEST 3 to assess the element of *is curious and aware*) ~~Did you wonder what student population the *Study Effectiveness Survey* would apply to? If so, what were your thoughts? If not, why not?~~
Changes: Based on the feedback from the expert reviewers that the item was limiting in eliciting respondents innate thought processes, the item was rewritten to be a more open-ended, ill-structured question to better elicit the element of *is curious and aware*.

Part IV: Extensions

The math department at the instructor's university heard that you helped in developing and using a measure of study effectiveness. They decide to hire you to help them investigate and evaluate their math students study habits.

12. (NEW ITEM to assess the elements of *develops a plan, applies previous knowledge to fit a new problem,* and *reasons with statistical models*) The math department is interested to see if there is a difference in the study habits between students in the two math course formats. They have the following question:

    > Is there a difference in the effectiveness of student study habits between students who enroll in face-to-face mathematics courses and those who enroll in online mathematics courses?

Using what you've learned in your statistic(s) courses, provide an brief outline of how the instructor should go about answering this question.

Changes: Based on the feedback from the expert reviewers about having no statistical inference within the assessment, the item was created to elicit respondents' knowledge about planning a study around a particular research question. Assessing the concept of statistical inference was incorporated into the item.

13. (NEW ITEM to assess the element of *reasons with statistical models*) The math department decided to implement a study skills intervention within their classes. They followed students over two years. Data from the survey were collected twice each semester, beginning and end, for a total of eight overall scores of study effectiveness. How might you summarize and present the data collected from the survey in a way that would be meaningful and useful to the math department?

Changes: Based on the feedback from the expert reviewers about having no statistical inference within the assessment, the item was created to assess respondent's knowledge of statistical models.

Study Effectiveness Survey

Difficulty of Topics:

1. How does the difficulty level of the topics in this course compare to other similar courses you have taken?

Much lower	Lower	About the same	Higher	Much higher

Prior Knowledge:

2. How does your prior knowledge about the topics in this course compare to other students in the class?

Much less	A little less	About the same	A little more	Much more

Scores on Assignments:

3. How do your assignment scores compare to other students in the class?

Much worse	A little worse	About the same	A little better	Much better

Grades:

4. How well do you think your current grade indicates how much you have learned in this course?

Not at all	Very little	Fair	To some extent	To a great extent

Time Spent Studying:

5. Approximately how much time each week do you spend studying to prepare for this class?

0-30 minutes	30-60 minutes	60-120 minutes	120-180 minutes	180 minutes or more

Quality of Time Studying:

6. How often are you distracted when you are studying for this class?

Never	Very rarely	Sometimes	Regularly	Very often

Skipping Parts:

7. How often do you skip parts of assignments or things that are important to do?

Never	Very rarely	Sometimes	Regularly	Very often

Read More than Once:

8. How often do you read the course material more than once?

Never	Very rarely	Sometimes	Regularly	Very often

Skim for Big Picture:

9. Before reading the course material, how often do you survey the chapter to develop a general idea of what the reading will be about?

Never	Very rarely	Sometimes	Regularly	Very often

Notetaking when Reading:

10. When reading the course material, how often do you take notes, highlight text, mark in the margins, or ask questions about the material?

Never	Very rarely	Sometimes	Regularly	Very often

Read and Not Comprehend:

11. When reading the course material, how often do you sacrifice comprehension for just getting through the pages assigned?

Never Very rarely Sometimes Regularly Very often

Synthesize the Readings:

12. When reading the course material, how often do you put what you've read into your own words or used your own examples to emphasize the main points?

Never Very rarely Sometimes Regularly Very often

Discuss with Others:

13. How often do you discuss the topics with other students in this course (e.g., try to simplify, summarize, or reorganize topics)?

Never Very rarely Sometimes Regularly Very often

Make Connections:

14. How often do you try to make connections between the topics learned in the course?

Never Very rarely Sometimes Regularly Very often

Appendix A4: MODEST (version 6)

Modeling To Elicit Statistical Thinking
(MODEST 6)[5]

INSTRUCTIONS

The purpose of this assessment is to evaluate how you think about a problem from a statistical point of view.

Directions:

1. Read the brief article to help you familiarize yourself with the problem scenario that you will be investigating.

2. Answer the questions related to solving the problem.
   - The questions are open-ended questions. Provide as much detail in your answers as possible so someone else can follow your thinking.
   - You will be evaluated based on how you describe your thought processes in your answers.
   - You may want to have writing materials available while you are solving the problem.

[5] Measuring Study Effectiveness (2013). In Pedagogy in Action: the SERC Portal for Educators. Retrieved July 15, 2014 from http://serc.carleton.edu/sp/library/mea/examples/example1.html.

Directions: Read the brief online news article to help you familiarize yourself with the problem scenario.

ONLINE NEWS ARTICLE

Effective Study Skills[6]

This web page on study skills, and what some may refer to as study tips, is designed to help you improve your learning and understanding, and ultimately your grades.

Here are some general techniques that seem to produce good results.

Strategies

When to study. The problem of when to study is critical. A good rule of thumb is that studying should be carried out only when you are rested, alert, and have planned for it.

Studying for lecture courses. If your study period is before the lecture class, be sure you have read all the assignments and made notes on what you don't understand. If the study period is after the lecture class, review the notes you took during class while the information is still fresh.

Thinking skills. If you're not a good thinker, start now by developing habits that make you ask yourself questions as you read. Talk to other students who you feel are good thinkers. Ask them what it is they do when they think critically or creatively.

The SQ3R method. SQ3R stands for Survey, Question, Read, Recite, Review.

> Survey - get the best overall picture of what you're going to study BEFORE you study it in any detail.
>
> Question - ask questions for learning.
>
> Read - When you read, read actively. Read to answer questions you have asked yourself or questions the instructor or author has asked.
>
> Recite - When you recite, you stop reading periodically to recall what you have read.
>
> Review - A review is a survey of what you have covered. It is a review of what you are supposed to accomplish, not what you are going to do.

Taking Lecture Notes

Surveying, Questioning, Listening. Your main job in taking lecture notes is to be a good listener. To be a good listener, you must learn to focus and concentrate on the main points of the lecture. Get them down, and then later reorganize them in your own words.

Reviewing and Revising. Begin the process by reviewing (and potentially revising) your notes right after a lecture. If you wait too long, you may discover that the notes just don't make sense.

A Final Word

The study skills presented here depend on one thing, and that is your willingness to WANT to improve and do well in school. You are the one who is responsible for your education, and effective study skills can help you. To that end, one last word of advice -- work smart, not hard.

[6] *Note.* Adapted from Kilzik, B. (2014, Sept. 30). Effective study skills. Retrieved on October 15, 2014 from http://www.adprima.com/studyout.htm.

Directions:
Answer the questions related to solving the problem.

- The questions are open-ended questions. Provide as much detail in your answers as possible so someone else can follow your thinking.
- You will be evaluated based on how you describe your thought processes in your answers.
- You may want to have writing materials available while you are solving the problem.

Task

Suppose you work as a survey design and analysis expert. An instructor of a college math course has hired you to create a survey ~~that instructors can use~~ to evaluate ~~their~~ her students' study habits. The goal is to have students take the survey about halfway through the semester. The results of the survey would be used to determine how effectively students are studying and to make improvements to their study habits, if needed.

PART I: Plan

Imagine that you were about to create the student study habits survey.

1. Think about the different factors that might affect student study habits (such as student characteristics, course characteristics, environment characteristics, etc.). Create a list of questions that you think should be on the survey to help you understand student study habits. To help you create the list, you may use the article provided, your own experience, and/or other external resources. For each question, briefly explain why student responses to that question would be useful when examining student study habits.

2. Do you think that the effect of the study habit factors on course grade would be the same for all students in the course? Explain why the factors would or would not have the same effect on course grade for all students.

3. In order to create the survey and use it to evaluate student study habits, what other kinds of information would you need to consider? Be sure to explain your reasoning.

Let's say the instructor of the college math course obtained a preexisting survey, called *Study Effectiveness Survey,* from a colleague and collected data from five of her students. See the handouts provided to familiarize yourself with the questions and format of the *Study Effectiveness Survey*.

PART II: Develop and Use a Measure of Study Effectiveness

Your first task is to develop an overall score of study effectiveness. This score will be a number that summarizes a student's overall study effectiveness and will help instructors judge the effectiveness of their students' study habits. You will be asked to try out your method on data from five students in the instructor's class. The questions below will guide you through this process.

4. When developing an overall score of study effectiveness, you will need to decide **how** each question on the *Study Effectiveness Survey* will or will not contribute to the overall score. Use the list of questions in the table below to describe **how** each question will or will not contribute to the overall score of study effectiveness. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the overall score.

	Question Content	
Question 1	**Difficulty of Material**	
Question 2	**Prior Knowledge**	
Question 3	**Current Grade**	
Question 4	**Grade as a Reflection of Learning**	
Question 5	**Time Spent Studying**	
Question 6	**Distracted when Studying**	
Question 7	**Skipping Parts**	
Question 8	**Read More than Once**	
Question 9	**Skim for Big Picture**	
Question 10	**Notetaking when Reading**	
Question 11	**Read and Not Comprehend**	
Question 12	**Synthesize the Readings**	
Question 13	**Discuss with Others**	
Question 14	**Make Connections**	

5. Using the answers you gave in question 4, describe how to compute an overall score of study effectiveness for a student.

6. (NEW ITEM to assess the statistical thinking element of *analyzes data*) Use your method se your method described in questions 4 and 5 to compute an overall score of study effectiveness for Al on the handout. Explain how you calculate the overall score for Al.

Changes: Based on the pilot students responses of not being able to articulate their method for computing an overall score of study effectiveness, this item was created to try to better understand their method.

7. Fill Al's result in the table below. ~~Using the instructor's data for her five students (see handout),~~ Then, repeat the process of using your method to calculate and report ~~each student's~~ an overall score of study effectiveness ~~using your method described earlier~~ for the four remaining students on the handout. Use the table below to help you with this process.

~~*Note*. You can go back at any time and adjust the method you described earlier. If you do this, **DON'T DELETE** your original answer to question 5. Describe the changes you would make in the space provided below the table.~~

Changes: Based on the pilot students responses of not adjusting their method, the note for this item was deleted. Additionally, the item was revised because of the addition of item 6 to MODEST 6.

	Question Content	Al	Barbara	Carl	Deborah	Ed
Question 1	**Difficulty of Material**					
Question 2	**Prior Knowledge**					
Question 3	**Current Grade**					
Question 4	**Grade as a Reflection of Learning**					
Question 5	**Time Spent Studying**					
Question 6	**Distracted when Studying**					
Question 7	**Skipping Parts**					
Question 8	**Read More than Once**					
Question 9	**Skim for Big Picture**					
Question 10	**Notetaking when Reading**					
Question 11	**Read and Not Comprehend**					
Question 12	**Synthesize the Readings**					
Question 13	**Discuss with Others**					
Question 14	**Make Connections**					
Score						

8. If two students have the same score of study effectiveness, does that mean they have the same study habits, according to the content on the survey? Explain your reasoning.

9. If the instructor ~~collected~~ used her survey to collect data from her whole class, describe the statistical measures or methods that you would use to ~~summarize how effective the instructor's students study habits are~~ provide a summary of the effectiveness of their study habits.

Changes: Item was revised to better clarify the task in the question.

PART III: Evaluation

Now that you have developed and used a measure of study effectiveness, your next task will be communicating the results to the instructor and evaluating and critiquing the process of measuring study effectiveness.

10. Write a brief report (~1-2 paragraphs) to the instructor that addresses:

    - how the overall score of study effectiveness was calculated using the *Study Effectiveness Survey*,
    - how to interpret an overall score of study effectiveness,
    - a summary of ~~how effective the study habits of the instructor's five students are~~ the overall scores for the five students,
    - the potential limitations of the *Study Effectiveness Survey*, and
    - ~~the potential limitations of~~ how well you think the overall score ~~of~~ measures study effectiveness.

Changes: Based on the pilot students responses that were lacking in communicating their understanding of the results, the item was modified to include an interpretation of the overall score and to clarify some of the (bulleted) details that go into the report.

11. What suggestions do you have to help the instructor revise the *Study Effectiveness Survey* or the overall score of study effectiveness*?*

Changes: Based on the pilot students responses providing suggestions of the score rather than the survey, the item was revised to have respondents provide suggestions for either the survey or the overall score of study effectiveness.

12. As a result of reading the article and working through the first ~~10~~ 11 questions, was there anything that you wondered about regarding the evaluation of student study habits? If so, what did you wonder about? If not, why not?

Part IV: Extensions

The math department at the instructor's university heard that you helped in developing and using a measure of study effectiveness. They decide to hire you to help them investigate and evaluate their math students study habits.

13. The math department is interested to see if there is a difference in the study habits between students in the two math course formats. They have the following question:

> Is there a difference in ~~the effectiveness of~~ student study habits between students who enroll in face-to-face mathematics courses and those who enroll in online mathematics courses?

Using what you've learned in your statistics course(s), provide a brief outline of how the instructor should go about answering this question.

A Sample of Five Student's Responses to the Study Effectiveness Survey

Question	Question Content	Al	Barbara	Carl	Deborah	Ed
1	**Difficulty of Material**	2 Easy	2 Easy	4 Hard	4 Hard	3 Moderate
2	**Prior Knowledge**	4 Much	2 A little	1 Not much	5 A great deal	4 Much
3	**Current Grade**	4 B	4 B	4 B	5 A	2 D
4	**Grades as a Reflection of Learning**	1 Not at all	2 Very little	4 To some extent	3 Fair	2 Very little
5	**Time Spent Studying**	3 60-120 minutes	1 0-30 minutes	4 120-180 minutes	5 180 minutes or more	1 0-30 minutes
6	**Distracted when Studying**	3 Sometimes	2 Very rarely	1 Never	3 Sometimes	4 Regularly
7	**Skipping Parts**	2 Very rarely	3 Sometimes	4 Regularly	3 Sometimes	2 Very rarely
8	**Read More than Once**	2 Very rarely	3 Sometimes	1 Never	2 Very rarely	2 Very rarely
9	**Skim for Big Picture**	4 Regularly	4 Regularly	1 Never	3 Sometimes	4 Regularly
10	**Notetaking when Reading**	1 Never	2 Very rarely	3 Sometimes	5 Very often	2 Very rarely
11	**Read and Not Comprehend**	3 Sometimes	4 Regularly	4 Regularly	4 Regularly	2 Very rarely
12	**Synthesize the Readings**	2 Very rarely	3 Sometimes	3 Sometimes	4 Regularly	2 Very rarely
13	**Discuss with Others**	2 Very rarely	1 Never	2 Very rarely	5 Very often	3 Sometimes
14	**Make Connections**	3 Sometimes	1 Never	2 Very rarely	5 Very often	3 Sometimes

Study Effectiveness Survey

Difficulty of Material:

1. How would you rate the difficulty of the course material?

1-Very easy	2-Easy	3-Moderate	4-Hard	5-Very hard

Prior Knowledge:

2. How much prior knowledge about the topics in this course did you have?

1-Not much	2-A little	3-Some	4-Much	5-A great deal

Current Grade:

3. What would you estimate your current grade would be in the course?

1- F	2- D	3- C	4- B	5- A

Grades as a Reflection of Learning:

4. How well do you think your current grade indicates how much you have learned in this course?

1 - Not at all	2-Very little	3 -Fair	4-To some extent	5-To a great extent

Time Spent Studying:

5. Approximately how much time each week do you spend studying to prepare for this class?

1- 0-30 minutes	2- 30-60 minutes	3- 60-120 minutes	4- 120-180 minutes	5- 180 minutes or more

Distracted when Studying:

6. How often are you distracted when you are studying for this class?

1-Never	2-Very rarely	3-Sometimes	4-Regularly	5-Very often

Skipping Parts:

7. How often do you skip parts of assignments or things that are important to do?

1-Never	2-Very rarely	3-Sometimes	4-Regularly	5-Very often

Read More than Once:

8. How often do you read the course material more than once?

1-Never	2-Very rarely	3-Sometimes	4-Regularly	5-Very often

Skim for Big Picture:

9. Before reading the course material, how often do you survey the chapter to develop a general idea of what the reading will be about?

1-Never	2-Very rarely	3-Sometimes	4-Regularly	5-Very often

Notetaking when Reading:

10. When reading the course material, how often do you take notes, highlight text, mark in the margins, or ask questions about the material?

1-Never	2-Very rarely	3-Sometimes	4-Regularly	5-Very often

Read and Not Comprehend:

11. When reading the course material, how often do you sacrifice comprehension for just getting through the pages assigned?

1-Never	2-Very rarely	3-Sometimes	4-Regularly	5-Very often

Synthesize the Readings:

12. When reading the course material, how often do you put what you've read into your own words or used your own examples to emphasize the main points?
    1-Never 2-Very rarely 3-Sometimes 4-Regularly 5-Very often

Discuss with Others:

13. How often do you discuss the topics with other students in this course (e.g., try to simplify, summarize, or reorganize topics)?
    1-Never 2-Very rarely 3-Sometimes 4-Regularly 5-Very often

Make Connections:

14. How often do you try to make connections between the topics learned in the course?
    1-Never 2-Very rarely 3-Sometimes 4-Regularly 5-Very often

Appendix B
Test Blueprints for MODEST

Appendix B1: Test Blueprint for Modeling To Elicit Statistical Thinking (MODEST 1) Assessment

COMPONENTS OF STATISTICAL THINKING	Item(s)
General Problem-Solving Characteristics	
• Creates a model *Description: Creates a model to help understand and predict a real-life situation.*	• Item 4
• Applies or adapts a previous problem to fit a new problem	• Item 5
• Seeks alternative explanations *Description: Seeks alternative explanations to help explain some response.*	• Item 11
Statistical Problem-Solving Processes	
• Develops a plan to carry out the analysis of the data, which includes: o Identifies types of information that is needed o Seeks information using pre-existing knowledge or an external source.	• Item 1 • Item 3
• Analyzes the data *Description: Fits and assesses a model to solve the problem.*	• Item 6
• Draws a conclusion *Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.*	• Item 8
Cognitive Processes of Statistical Problem-Solving	
• Considers variation *Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables*	• Item 2 • Item 4 • Item 7
• Transforms the raw data into an aggregate form (e.g., graphs, numerical summaries)	• Item 6
• Reasons with statistical models *Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics.*	• Item 9
• Integrates the statistical and contextual information	• Item 8
Individual Dispositions	

• Is curious and aware *Description: Is curious by asking Items such as, "is this something that happens more generally?"*	• Item 12
• Is open to new ideas *Description: Considers new ideas or information that conflict with own knowledge or assumption.*	• Item 3
• Is innovative	• Item 4
• Is skeptical (e.g., is this conclusion justified?)	• Item 10
• Is logical (e.g., constructs a logical argument)	• Item 1

Appendix B2: Test Blueprint for Modeling To Elicit Statistical Thinking (MODEST 4) Assessment

COMPONENTS OF STATISTICAL THINKING	Item(s)
General Problem-Solving Characteristics	
• Creates a model *Description: Creates a model to help understand and predict a real-life situation.*	• Item 5
• Applies previous knowledge or adapts a previous problem to fit a new problem	• Item 8 • Item 12
• Seeks alternative explanations *Description: Seeks alternative explanations to help explain some response.*	• Item 10
Statistical Problem-Solving Processes	
• Develops a plan for collection or analysis of the data, which includes: o Identifying types of information that is needed o Seeking information using pre-existing knowledge or an external source.	• Item 1 • Item 3 • Item 12
• Analyzes the data *Description: Fits and assesses a model to solve the problem.*	• Item 6
• Draws a conclusion *Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.*	• Item 9
Cognitive Processes of Statistical Problem-Solving	
• Considers variation *Description: Includes:* o *Explaining variation among variables or cases.* o *Looking for sources of variability by examining patterns in the variables or relationships between variables.* o *Considering measurement error.*	• Item 2 • Item 4 • Item 7
• Reasons with statistical models *Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics.*	• Item 8 • Item 9 • Item 12 • Item 13
• Integrates the statistical and contextual information	• Item 9
Individual Dispositions	
• Is curious *Description: Is curious by asking Items such as, "is this something that happens more generally?"*	• Item 11

• Is open to new ideas *Description: Considers new ideas or information*	• Item 3
• Is innovative	• Item 4
• Is skeptical (e.g., is this conclusion justified?)	• Item 9
• Is logical (e.g., constructs a logical argument)	• Item 1

Appendix B3: Test Blueprint for Modeling To Elicit Statistical Thinking (MODEST 6) Assessment

COMPONENTS OF STATISTICAL THINKING	Item(s)
General Problem-Solving Characteristics	
• Creates a model *Description: Creates a model to help understand and predict a real-life situation.*	• Item 4 • Item 5
• Applies previous knowledge or adapts a previous problem to fit a new problem	• Item 9 • Item 13
• Seeks alternative explanations *Description: Seeks alternative explanations to help explain some response.*	• Item 11
Statistical Problem-Solving Processes	
• Develops a plan for collection or analysis of the data, which includes: o Identifying types of information that is needed o Seeking information using pre-existing knowledge or an external source.	• Item 1 • Item 3 • Item 13
• Analyzes the data *Description: Fits and assesses a model to solve the problem.*	• Item 6 • Item 7
• Draws a conclusion *Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.*	• Item 10
Cognitive Processes of Statistical Problem-Solving	
• Considers variation *Description: Includes:* o *Explaining variation among variables or cases.* o *Looking for sources of variability by examining patterns in the variables or relationships between variables.* o *Considering measurement error.*	• Item 2 • Item 8
• Reasons with statistical models *Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics.*	• Item 9 • Item 10 • Item 13
• Integrates the statistical and contextual information	• Item 10
• Recognizes the need for data	• Item 13
Individual Dispositions	
• Is curious	• Item 12

Description: Is curious by asking Items such as, "is this something that happens more generally?"	
• Is open to new ideas	-
Description: Considers new ideas or information	
• Is innovative	• Item 4
• Is skeptical/critical (e.g., is this conclusion justified?)	• Item 10
• Is logical (e.g., constructs a logical argument)	• Item 1

Appendix B4: FINAL Test Blueprint for Modeling To Elicit Statistical Thinking (MODEST 6) Assessment

Components Of Statistical Thinking	Original Item(s)	Final Item(s)
General Problem-Solving Characteristics		
• Creates a model o Produces a conceptual model o Translates the conceptual model into a statistical model o Quality of the model *Description: Creates a model to help understand and predict a real-life situation.*	• Item 4 • Item 5	• Item 4: Produces a conceptual model • Item 5: o Translates the conceptual model into a statistical model o Quality of the model
• Applies previous knowledge or adapts a previous problem to fit a new problem • Seeks alternative explanations *Description: Seeks alternative explanations to help explain some response.*	• Item 9 • Item 13 • Item 11	-
Statistical Problem-Solving Processes		
• Develops a [reasonable] plan for collection or analysis of the data, which includes: o Identifying types of information that is needed o Seeking information using pre-existing knowledge or an external source.	• Item 1 • Item 3 • Item 13	• Item 1 • Item 3 • Item 13
• Analyzes the data *Description: Fits and assesses a model to solve the problem.*	• Item 6 • Item 7	• Item 7
• Draws a conclusion *Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.*	• Item 10	• Item 10: Interprets findings.
Cognitive Processes of Statistical Problem-		

Solving		
• Considers variation *Description: Includes:* o *Explaining variation among variables or cases.* o *Looking for sources of variability by examining patterns in the variables or relationships between variables.* o *Considering measurement error.*	• Item 2 • Item 8	• Item 2 • Item 8
• Reasons with statistical models *Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics.*	• Item 9 • Item 10 • Item 13	• Item 9 • Item 10 • Item 13
• Integrates the statistical and contextual information	• Item 10	-
• Recognizes the need for data	• Item 13	• Item 13
Individual Dispositions		
• Is curious *Description: Is curious by asking Items such as, "is this something that happens more generally?"*	• Item 12	• Item 12
• Is open to new ideas *Description: Considers new ideas or information*	-	-
• Is innovative/creative	• Item 4	-
• Is skeptical (e.g., is this conclusion justified?)	• Item 10	• Item 10 • Item 11
• Is logical (e.g., constructs a logical argument)	• Item 1	-

Appendix C
Expert Reviewer Materials

Appendix C1: Invitation Email to Expert Reviewers

Dear Professor XXX,

I am writing to ask for your assistance as part of the review process for developing the *Modeling To Elicit Statistical Thinking* (MODEST) assessment for the introductory statistics course. The development of this instrument is part of my doctoral dissertation in Statistics Education at the University of Minnesota, under the supervision of my co-advisors Joan Garfield and Andrew Zieffler.

I am requesting your help with this project because your expertise, knowledge, and background suggests you would be a good reviewer of this assessment. If you agree to review MODEST, you will be asked to provide feedback on a preliminary version of the assessment. Your feedback will be invaluable for providing evidence of validity and helping me to further refine MODEST.

Here is a brief background on the development of MODEST:

MODEST uses an ill-structured problem designed to elicit students' statistical thinking. Adapted from a model-eliciting activity (Study Effectiveness MEA; "Measuring Study Effectiveness", 2009), MODEST was written to be an individual assessment utilizing constructed-response items. The test blueprint used during item development was based on integrating elements of statistical thinking identified in Wild and Pfannkuch's (1999) empirical framework and characteristics of *expert thinking* from the expert-novice literature.

To aid in the review process, I have also attached the test blueprint and an evaluation form. The instructions for the review process are located on the first page of the evaluation form. Ideally, I would like to receive your review by September 10, 2014. Please let me know whether you can provide this review by replying to this email (free0312@umn.edu). Feel free to contact me with any questions.

I hope you will agree to participate as an expert reviewer of the preliminary version of MODEST. Thank you for your time and consideration.

Sincerely,

Laura Le
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

References:

Measuring Study Effectiveness (2009). In P*edagogy in Action: the SERC Portal for Educators*. Retrieved June 7, 2014 from http://serc.carleton.edu/sp/library/mea/examples/example1.html.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223-248.

Appendix C2: MODEST Assessment provided to Expert Reviewers

Modeling To Elicit Statistical Thinking
(MODEST 3)[7]

INSTRUCTIONS

The purpose of this assessment is to evaluate how you think about a problem from a statistical point of view.

Directions:

1. Read the brief article to help you familiarize yourself with the problem scenario that you will be investigating.

2. Answer the questions related to solving the problem.
   - The questions are open-ended questions. Provide as much detail in your answers as possible so someone else can follow your thinking.
   - You will be evaluated based on how you describe your thought processes in your answers.
   - You may want to have writing materials available while you are solving the problem.

[7] Adapted from an activity created by Richard Lesh.

ONLINE NEWS ARTICLE

Effective study skills for getting best exam results[8]

As students prepare to return to school, it is important to get off to a good September start by ensuring that students have the necessary skills required to maximize study time. Students may participate in a study skills seminar that is especially designed to help students organize their study and homework as they return to school. Study skills seminars typically focus on a number of key topics:

Maximizing Class Time-Taking and Making Notes

Students learn the importance of creating an effective class and learn how to organize notes. Summary notes play a vital role during exam revision.

Devising Realistic Study Plans

Students learn how to create effective, realistic study plans for their homework and revision sessions.

Dealing with Homework

Since homework is an integral part of any successful exam result, students are advised on how to handle homework as part of an effective study schedule.

Memory Techniques

Students are given multiple memory techniques to help in the retention and recollection of information during exams.

Exam Techniques

Students are introduced to invaluable exam-taking techniques.

Study and Staying Healthy

Students learn the importance of studying and staying healthy while preparing for exams.

All in all, study skills seminars get students to focus on successful methods that help them organize their studies and improve their grades in school.

[8] *Note.* Adapted from Effective study skills for getting best exam results (2009, Aug 6). Retrieved on July 3, 2014 from http://www.limerickpost.ie/2009/08/06/effective-study-skills-for-getting-best-exam-results/.

Task

Suppose you work as a survey design and analysis expert. Tutors for a study skills seminar have hired you to create a survey that helps students self-assess their own study habits for a course. The goal is to have students take the survey about halfway through the semester. The results of the survey would be used to determine how effectively students are studying and to make improvements to their study habits, if needed.

PART I: Plan

Imagine that you were about to create the survey for the tutors.

1. List student or class characteristics that you think would be useful for determining the effectiveness of students' study habits. Use the article provided, your own experience, and/or other external resources for creating your list. For each characteristic, briefly explain why you would include it on the survey.

2. Are the characteristics you listed of equal value with respect to helping a student be an effective studier? Explain your reasoning.

3. Would asking the tutors questions about their knowledge of students' study habits be useful when creating the survey? Why or why not?

[NEW ONLINE PAGE]

Let's say you created the survey and called it the *Study Effectiveness Survey*. See the following link to familiarize yourself with the questions and format of the *Study Effectiveness Survey*:

[*Note. The online assessment will have a link to the survey right here. For the review process, see page 8-9]*

PART II: Develop and Test a Measure of Study Effectiveness

Your first task is to develop a summary score of study effectiveness that will be used on the students' responses to the survey. This score will help the students judge how effective their study habits are. Be sure to refer to the survey and the information in the table below to help you come up with your score.

A Sample of Five Student's Responses to the Study Effectiveness Survey

Question	Question Content	Al	Barbara	Carl	Deborah	Ed
1	**Difficulty of Topics**	Lower	Lower	Higher	Higher	About the same
2	**Prior Knowledge**	A little more	A little less	Much less	Much more	A little more
3	**Scores on Assignments**	A little better	A little better	A little better	Much better	A little worse
4	**Grades**	Not at all	Very little	To some extent	Fair	Very little
5	**Time Spent Studying**	60-120 minutes	0-30 minutes	120-180 minutes	180 minutes or more	0-30 minutes
6	**Quality of Time Studying**	Sometimes	Very rarely	Never	Sometimes	Regularly
7	**Skipping Parts**	Very rarely	Sometimes	Regularly	Sometimes	Very rarely
8	**Read More than Once**	Very rarely	Sometimes	Never	Very rarely	Very rarely
9	**Skim for Big Picture**	Regularly	Regularly	Never	Sometimes	Regularly
10	**Notetaking when Reading**	Never	Very rarely	Sometimes	Very often	Very rarely
11	**Read and Not Comprehend**	Sometimes	Regularly	Regularly	Regularly	Very rarely
12	**Synthesize the Readings**	Very rarely	Sometimes	Sometimes	Regularly	Very rarely
13	**Discuss with Others**	Very rarely	Never	Very rarely	Very often	Sometimes

14	**Make Connections**	Sometimes	Never	Veryrarely	Very often	Sometimes

4. To aid in the process of developing a summary score of study effectiveness, fill in the table below describing how each question will contribute to the overall score of study effectiveness. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the score.

Question 1	
Question 2	
Question 3	
Question 4	
Question 5	
Question 6	
Question 7	
Question 8	
Question 9	
Question 10	
Question 11	
Question 12	
Question 13	
Question 14	
Score	

5. Report how to produce your summary score of study effectiveness given any student's responses to the survey. Be sure to include any formulas, procedures, or rules on which your score of study effectiveness is based.

6. For each of the five students in the table, calculate and report their score of study effectiveness using your method described earlier. Also, place these students in rank order according to their study effectiveness. A table is provided below to help you with this process.

	Al	Barbara	Carl	Deborah	Ed
Question 1					
Question 2					
Question 3					
Question 4					
Question 5					
Question 6					
Question 7					

Question 8					
Question 9					
Question 10					
Question 11					
Question 12					
Question 13					
Question 14					
Score					

7. If two students have the same score of study effectiveness, does that mean they have the same study habits, according to the content on the survey? Explain your reasoning.

[NEW ONLINE PAGE]

PART III: Evaluation

8. Write a report to the tutors of the study skills seminar that describes:

    - how to calculate your summary score of study effectiveness using the *Study Effectiveness Survey,* and
    - how to interpret the summary score so individual students can judge how effective their study habits are.

9. Do you think it would be useful to provide students with a summary graph of their results on the survey? If so, what type of graph would you present and how would this be useful to the students? If not, why not?

10. What concerns or reservations would you have about using the *Study Effectiveness Survey* or your score of study effectiveness within a classroom? Explain.

11. Are there other student or classroom characteristics you would consider examining to help refine your survey of effective study habits? Explain your reasoning.

12. Did you wonder what student population the *Study Effectiveness Survey* would apply to? If so, what were your thoughts? If not, why not?

Study Effectiveness Survey

Difficulty of Topics:

1. How does the difficulty level of the topics in this course compare to other similar courses you have taken?

Much lower	Lower	About the same	Higher	Much higher

Prior Knowledge:

2. How does your prior knowledge about the topics in this course compare to other students in the class?

Much less	A little less	About the same	A little more	Much more

Scores on Assignments:

3. How do your assignment scores compare to other students in the class?

Much worse	A little worse	About the same	A little better	Much better

Grades:

4. How well do you think your current grade indicates how much you have learned in this course?

Not at all	Very little	Fair	To some extent	To a great extent

Time Spent Studying:

5. Approximately how much time each week do you spend studying to prepare for this class?

0-30 minutes	30-60 minutes	60-120 minutes	120-180 minutes	180 minutes or more

Quality of Time Studying:

6. How often are you distracted when you are studying for this class?

Never	Very rarely	Sometimes	Regularly	Very often

Skipping Parts:

7. How often do you skip parts of assignments or things that are important to do?

Never	Very rarely	Sometimes	Regularly	Very often

Read More than Once:

8. How often do you read the course material more than once?

Never	Very rarely	Sometimes	Regularly	Very often

Skim for Big Picture:

9. Before reading the course material, how often do you survey the chapter to develop a general idea of what the reading will be about?

Never	Very rarely	Sometimes	Regularly	Very often

Notetaking when Reading:

10. When reading the course material, how often do you take notes, highlight text, mark in the margins, or ask questions about the material?

Never	Very rarely	Sometimes	Regularly	Very often

Read and Not Comprehend:

11. When reading the course material, how often do you sacrifice comprehension for just getting through the pages assigned?

Never	Very rarely	Sometimes	Regularly	Very often

Synthesize the Readings:

12. When reading the course material, how often do you put what you've read into your own words or used your own examples to emphasize the main points?

Never Very rarely Sometimes Regularly Very often

Discuss with Others:

13. How often do you discuss the topics with other students in this course (e.g., try to simplify, summarize, or reorganize topics)?

Never Very rarely Sometimes Regularly Very often

Make Connections:

14. How often do you try to make connections between the topics learned in the course?

Never Very rarely Sometimes Regularly Very often

Appendix C3: Test Blueprint provided to Expert Reviewers

Test Blueprint for
Modeling To Elicit Statistical Thinking (MODEST 3) Assessment

COMPONENTS OF STATISTICAL THINKING	Item(s)
General Problem-Solving Characteristics	
• Creates a model *Description: Creates a model to help understand and predict a real-life situation.*	• Question 5
• Applies previous knowledge or adapts a previous problem to fit a new problem	• Question 4
• Seeks alternative explanations *Description: Seeks alternative explanations to help explain some response.*	• Question 11
Statistical Problem-Solving Processes	
• Develops a plan to carry out the analysis of the data, which includes: o Identifying types of information that is needed o Seeking information using pre-existing knowledge or an external source.	• Question 1 • Question 3
• Analyzes the data *Description: Fits and assesses a model to solve the problem.*	• Question 6
• Draws a conclusion *Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.*	• Question 8
Cognitive Processes of Statistical Problem-Solving	
• Considers variation *Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables*	• Question 2 • Question 4 • Question 7
• Transforms the raw data into an aggregate form (e.g., graphs, numerical summaries)	• Question 6
• Reasons with statistical models *Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics.*	• Question 9
• Integrates the statistical and contextual information	• Question 8
Individual Dispositions	
• Is curious and aware *Description: Is curious by asking questions such as, "is this something that happens more generally?"*	• Question 12
• Is open to new ideas *Description: Considers new ideas or information that conflict with own knowledge or assumption.*	• Question 3

• Is innovative • Is skeptical (e.g., is this conclusion justified?) • Is logical (e.g., constructs a logical argument)	• Question 5 • Question 10 • Question 1 • Question 4

Appendix C4: Evaluation Form provided to Expert Reviewers

Evaluation Form for MODEST

Instructions for the review process:

1. Read through the MODEST assessment.
2. Look over the test blueprint that was used to develop the assessment.
3. Complete Parts 1 and 2 of this evaluation form.
4. Email this evaluation form back to me at free0312@umn.edu.

Part 1. Evaluation at the item level.

For each assessment item, rate the extent to which you agree or disagree that the item measures the intended statistical thinking element, as shown below:

1. List student or class characteristics that you think would be useful for determining the effectiveness of students' study habits. Use the article provided, your own experience, and/or other external resources for creating your list. For each characteristic, briefly explain why you would include it on the survey.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Statistical Problem-Solving Behaviors – Develops a plan to carry out the analysis of the data, which includes: • Identifying types of information that is needed • Seeking information using pre-existing knowledge or an external source.			
Comments (optional):			
Individual Dispositions – Is logical (e.g., constructs a logical argument)			
Comments (optional):			

2. Are the characteristics you listed of equal value with respect to helping a student be an effective studier? Explain your reasoning.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Considers variation Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables			
Comments (optional):			

3. Would asking the tutors questions about their knowledge of students' study habits be useful when creating the survey? Why or why not?

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Statistical Problem-Solving Behaviors – Develops a plan to carry out the analysis of the data, which includes: • Identifying types of information that is needed • Seeking information using pre-existing knowledge or an external source.			
Comments (optional):			

Individual Dispositions – Is open to new ideas Description: Considers new ideas or information that conflict with own knowledge or assumption.			
Comments (optional):			

4. To aid in the process of developing a summary score of study effectiveness, fill in the table below describing how each question will contribute to the overall score of study effectiveness. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the score.

Question 1	
Question 2	
Question 3	
Question 4	
Question 5	
Question 6	
Question 7	
Question 8	
Question 9	
Question 10	
Question 11	
Question 12	
Question 13	
Question 14	
Score	

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
General Problem-Solving Characteristics – Applies previous knowledge or adapts a previous problem to a new situation			
Comments (optional):			

Cognitive Processes of Statistical Problem-Solving – Considers variation Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables			
Comments (optional):			
Individual Dispositions – Is logical (e.g., constructs a logical argument)			
Comments (optional):			

5. Report how to produce your summary score of study effectiveness given any student's responses to the survey. Be sure to include any formulas, procedures, or rules on which your score of study effectiveness is based.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
General Problem-Solving Characteristics – Creates a model Description: Creates a model to help understand and predict a real-life situation.			
Comments (optional):			
Individual Dispositions – Is innovative			

Comments (optional):			

6. For each of the five students in the table, calculate and report their score of study effectiveness using your method described earlier. Also, place these students in rank order according to their study effectiveness. A table is provided below to help you with this process.

	Al	Barbara	Carl	Deborah	Ed
Question 1					
Question 2					
Question 3					
Question 4					
Question 5					
Question 6					
Question 7					
Question 8					
Question 9					
Question 10					
Question 11					
Question 12					
Question 13					
Question 14					
Score					

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
*Cognitive Processes of Statistical Problem-Solv*ing – Transforms the raw data into an aggregate form (e.g., graphs, numerical summaries)			
Comments (optional):			

Statistical Problem-Solving Behaviors – Analyzes the data Description: Fits and assesses a model to solve the problem.			
Comments (optional):			

7. If two students have the same score of study effectiveness, does that mean they have the same study habits, according to the content on the survey? Explain your reasoning.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Considers variation Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables			
Comments (optional):			

8. Write a report to the tutors of the study skills seminar that describes:

    - how to calculate your summary score of study effectiveness using the *Study Effectiveness Survey,* and
    - how to interpret the summary score so individual students can judge how effective their study habits are.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with	Disagree

		reservations	
Statistical Problem-Solving Behaviors – Draws a conclusion Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.			
Comments (optional):			
Cognitive Processes of Statistical Problem-Solving – Integrates the statistical and contextual information			
Comments (optional):			

9. Do you think it would be useful to provide students with a summary graph of their results on the survey? If so, what type of graph would you present and how would this be useful to the students? If not, why not?

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Reasons with statistical models Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics			
Comments (optional):			

10. What concerns or reservations would you have about using the *Study Effectiveness Survey* or your score of study effectiveness within a classroom? Explain.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Individual Dispositions – Is skeptical (e.g., is this conclusion justified?)			
Comments (optional):			

11. Are there other student or classroom characteristics you would consider examining to help refine your survey of effective study habits? Explain your reasoning.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
General Problem-Solving Characteristics – Seeks alternative explanations Description: Seeks alternative explanations to help explain some response.			
Comments (optional):			

12. Did you wonder what student population the *Study Effectiveness Survey* would apply to? If so, what were your thoughts? If not, why not?

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Individual Dispositions– Is curious and aware Description: Considers new ideas or information that conflict with own knowledge or assumption.			
Comments (optional):			

PART 2. EVALUATION AT THE ASSESSMENT LEVEL.

How much do you agree or disagree with the following statement?

	Agree	Agree, but with reservations	Disagree
Overall, the assessment appears to measure statistical thinking, based off of the test blueprint.			
Comments (optional):			

Please add any suggestions you have for improving MODEST.

Thank you for helping develop MODEST!

Appendix D
Results of Feedback from the Expert Reviewers

Appendix D1: Expert Reviewer Agreement Ratings

Table D1

Frequency of reviewers' responses to the items in MODEST 3

Item #	Statistical Thinking Element	Agree	Agree, but with Reservations	Disagree	NA
Item 1					
	Statistical Problem-Solving Behaviors – Develops a plan to carry out the analysis of the data.	2	3	0	0
	Individual Dispositions – Is logical (e.g., constructs a logical argument).	2	3	0	0
Item 2					
	Cognitive Processes of Statistical Problem-Solving – Considers variation.	0	2	3	0
Item 3					
	Statistical Problem-Solving Behaviors – Develops a plan to carry out the analysis of the data.	0	4	1	0
	Individual Dispositions – Is open to new ideas.	1	2	2	0
Item 4					
	General Problem-Solving Characteristics – Applies previous knowledge or adapts a previous problem to a new situation.	3	1	0	1
	Cognitive Processes of Statistical Problem-Solving – Considers variation.	2	2	1	0
	Individual Dispositions – Is logical (e.g., constructs a logical argument).	5	0	0	0
Item 5					

	General Problem-Solving Characteristics – Creates a model.	3	1	1	0
	Individual Dispositions – Is innovative.	2	2	0	1
Item 6					
	*Cognitive Processes of Statistical Problem-Solv*ing – Transforms the raw data into an aggregate form (e.g., graphs, numerical summaries).	3	2	0	0
	Statistical Problem-Solving Behaviors – Analyzes the data.	2	0	2	1
Item 7					
	Cognitive Processes of Statistical Problem-Solving – Considers variation.	2	2	1	0
Item 8					
	Statistical Problem-Solving Behaviors – Draws a conclusion.	2	2	1	0
	Cognitive Processes of Statistical Problem-Solving – Integrates the statistical and contextual information.	4	1	0	0
Item 9					
	Cognitive Processes of Statistical Problem-Solving – Reasons with statistical models.	3	2	0	0
Item 10					
	Individual Dispositions – Is skeptical (e.g., is this conclusion justified?).	2	3	0	0
Item 11					
	General Problem-Solving Characteristics – Seeks alternative explanations.	1	3	1	0
Item 12					

Individual Dispositions– Is curious and aware.	2	2	1	0

Appendix D2: Expert Reviewer Comments regarding Items and Subsequent Changes Made to MODEST

1. List student or class characteristics that you think would be useful for determining the effectiveness of students' study habits. Use the article provided, your own experience, and/or other external resources for creating your list. For each characteristic, briefly explain why you would include it on the survey.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Statistical Problem-Solving Behaviors – Develops a plan to carry out the analysis of the data, which includes: • Identifying types of information that are needed • Seeking information using pre-existing knowledge or an external source.	Reviewer #3 Reviewer #4	Reviewer #1 Reviewer #2 Reviewer #5	

Comments:

Reviewer #1:
The question is well aligned, however, the inclusion of the preceding article may prime student responses. You may not get a reliable indication of how well a student can think within the context. Could it be completely open-ended by removing the news article? It also ties into the consideration of variation, as students need think about factors that explain variability in effective study habits. It also ties into integrating the statistical and contextual. As you will find, there is often little meaningful distinction between elements of thinking. Every statistical thinking question is in some way multidimensional. If it wasn't, it's probably not a good thinking question.

Reviewer #2:
I just wouldn't call this "analysis of the data" but a plan for collection of the data. Give a target number of characteristics to list? It's not totally clear to me how you would differentiate strong vs. weak answers. Partly I'm not sure whether you are getting at "inputs" or "outputs." You want to know how they might measure effectiveness? You mean characteristics of good study habits or good things to measure to help the students self-assess? When you say "include it on the survey" do you mean what questions they could ask to assess effectiveness of student habits? Maybe just phrase it like that rather than in terms of characteristics which might throw students off?

Reviewer #4:
I found this question confusing. How can a class or student characteristic DETERMINE the effectiveness of a students' study habits? Are characteristics that you are thinking about here include: gender, whether they are repeated the course, year in college, major, age, discipline of the course, whether course meets 3 times a week or once a week? etc? That is what initially comes ot mind….but I THINK you are also interested in getting students to think about what to measure about study habits as well (how long they study, do they write notes outside of the class? Do they study in groups?). Instead do you want to ask what class or student characteristics might be useful to know when examining students' study habits.

Reviewer #5:
I think at this level to rely solely on pre-existing knowledge is not a good stats problem-solving behavior. Should be Seeking information using pre-existing knowledge AND an external source

---

Changes:

- Description of the *develops a plan* element was modified (Reviewer #2 & Reviewer #5)
- Item was revised to clarify what is being asked (Reviewer #2 & Reviewer #4).

---

2. Are the characteristics you listed of equal value with respect to helping a student be an effective studier? Explain your reasoning.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Considers variation Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables		Reviewer #1 Reviewer #5	Reviewer #2 Reviewer #3 Reviewer #4

Comments:

Reviewer #1:
This question asks whether the indicators from Q1 can be considered equal. For example, is note-taking equally important as homework? The answer is probably, no. This ties into the concept of explaining variation. I agree that it share a link, but perhaps moves a little away from the "big" concepts of variation, for example measurement error (induced), between groups variation, within subjects variation and sampling variability.

Reviewer #2:
I agree it's important to think that some variables may be more relevant than others, but I really don't think this has to do with variation. I guess you could argue it has to do with trying to judge strength of association, but I'm not sure you are getting at both explanatory and response variables in question 1. It also seems to change focus a bit here, am I helping a student be a more effective studier or am I helping a student self-evaluate his or her study habits?

Reviewer #3:
No data are provided for analysis in this item. It appears to be a question requiring a priori judgment about weighting rather than formulating an appropriate model that is grounded in the data.

Reviewer #4:
It is not clear to me how this item will elicit students reasoning about variation. At first glance, I think this question is asking me if some characteristics are more important than others. So is their gender more important? Does it matter if they are repeating the course or not? Does it matter if the class meets 3 times a week or once a week? But if you fix the first question to get them to think about what they should be measuring related to students study habits, then this question (if made more clear), may eleicit more thinking about variation.

Reviewer #5:
You need a wider description for considers variation, as there is no data at this stage. I think you could have something about Managing variation.

Changes:

- Item was rewritten to try to better elicit the element of *considers variation* (Reviewer #1, Reviewer #2, Reviewer #3, & Reviewer #4)
- Description of the *considers variation* element was modified (Reviewer #3 & Reviewer #5)

---

3. Would asking the tutors questions about their knowledge of students' study habits be useful when creating the survey? Why or why not?

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Statistical Problem-Solving Behaviors – Develops a plan to carry out the analysis of the data, which includes: • Identifying types of information that is needed • Seeking information using pre-existing knowledge or an external source.		Reviewer #1 Reviewer #2 Reviewer #3 Reviewer #5	Reviewer #4

Comments:

Reviewer #1
Yes, as students are required to compare and select appropriate sources of information. This question also considers variation. How do tutor accounts of effective study habits vary from students?

Reviewer #2
Again I would differentiate data collection from data analysis
I'm also again a little confused on how strong and weak answers will be judged. Or are you looking more at the justification?

Reviewer #3
It might be productive here to leave open the possibility of drawing on any additional source of information and then letting the respondent identify possible sources.

Reviewer #4
The question seems to be about seeking information for planning instrument development. If you want the question to measure their ability to develop a plan for analysis of data, then the question should be more direct at asking if a discussion with the tutors could inform the plan for how to analyze data and present findings to the tutors.

Reviewer #5
To design a survey it would seem that the designer would need to be using relevant and appropriate measures for the study skill course that the students were given. That is, the survey is related to the actual course. So your description should incorporate the notion of relevancy.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
	Reviewer #2	Reviewer #4 Reviewer #5	Reviewer #1 Reviewer #3
Individual Dispositions – Is open to new ideas Description: Considers new ideas or information that conflict with own knowledge or assumption.			

Comments:

Reviewer #1
Considering the definition, I don't think this question qualifies for "conflicting with one's own knowledge". Alternative approaches do not necessarily conflict with one's primary approach. However, if "conflict" is not a qualifier for the definition, then, yes, this question does relate to considering alternative approaches.

Reviewer #2
So you can see if they agree to hear these perspectives but not totally clear they would do it on their own (vs. a question asking what other steps they might take or even what additional information they might seek)

Reviewer #3
There does not seem to be anything in the item that conflicts with common assumptions.

Reviewer #4
I am not convinced that this question will elicit a response that will give you insight into whether someone has this disposition. I would like to know your anticipated responses that you think would give insight into this disposition.

Reviewer #5
From this question I am not sure whether you could ascertain whether students are considering information that conflicts with their own knowledge or assumptions.

Changes:

- Description of the *is open to new ideas* element was modified (Reviewer #1, Reviewer #3, & Reviewer #5)
- Item was rewritten to be a more open-ended, ill-structured question (Reviewer #2 & Reviewer #3).

4. To aid in the process of developing a summary score of study effectiveness, fill in the table below describing how each question will contribute to the overall score of study effectiveness. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the score.

Question 1	
Question 2	
Question 3	
Question 4	
Question 5	
Question 6	
Question 7	
Question 8	
Question 9	
Question 10	
Question 11	
Question 12	
Question 13	
Question 14	
Score	

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
General Problem-Solving Characteristics – Applies previous knowledge or adapts a previous problem to a new situation	Reviewer #1 Reviewer #3 Reviewer #5	Reviewer #4	

Comments:

Reviewer #1

Assuming the student covers similar or related problems in their course, this question would be an opportunity for the students to apply or adapt previous knowledge.

Reviewer #2
Ok if the previous knowledge is what you told them earlier, that some questions might be more important than others? Is this closely enough related to statistical thinking to be relevant?

Reviewer #3
The part about applying previous knowledge seems applicable; the part about adapting a previous problem is not.

Reviewer #4
Not sure how you will know the prior knowledge they use or if they are adapting a strategy from a previous problem they have done. Maybe you should ask something more direct in the question to elicit this.

Reviewer #5
I assume students have done a similar problem.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Considers variation Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables	Reviewer #4 Reviewer #5	Reviewer #1 Reviewer #3	Reviewer #2

Comments:

Reviewer #1
The student has to consider numerous variations of scoring to determine an approach that best suits the problem and reflects the construct under investigation. The student needs to consider relationships between indictors and weight the appropriate items to best "model" effective study habits. I believe this items also taps strongly into transnumeration, or how numbers can be used to gain insight into a problem, e.g. scoring study habits numerically.

Reviewer #2

I'm not sure how this elicits enough on variation. I think the statistical issue is more about relevance and how to code the items. It might be more useful to not always put the "neutral" response in the middle and then see whether or not the student adjusts to that?

Reviewer #3
The same observation as given for an earlier item applies here as well: Respondents are not given any data to analyze, and data analysis seems to be an important type of reasoning to assess.

Reviewer #5
Again there is no data so I think managing sources of variability should be part of the description.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Individual Dispositions – Is logical (e.g., constructs a logical argument)	Reviewer #1 Reviewer #2 Reviewer #3 Reviewer #4 Reviewer #5		

Comments:

Reviewer #1
Yes, the student is required to provide a rationale for their scoring criteria. This also relates to integrating the statistical and contextual.

5. Report how to produce your summary score of study effectiveness given any student's responses to the survey. Be sure to include any formulas, procedures, or rules on which your score of study effectiveness is based.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree

General Problem-Solving Characteristics – Creates a model Description: Creates a model to help understand and predict a real-life situation.	Reviewer #1 Reviewer #4 Reviewer #5	Reviewer #3	Reviewer #2

Comments:

Reviewer #1
Yes, this is an example of applying a model to solve a problem.

Reviewer #2
Seems like questions 4 and 5 need to be combined? Are all formulas models?

Reviewer #3
A statistical model would generally be grounded in data, but in this case, no data are provided.

Reviewer #5
I think the description should include something about justifying the model.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Individual Dispositions – Is innovative	Reviewer #1 Reviewer #4	Reviewer #2 Reviewer #3	

Comments:

Reviewer #1
As the student has little to go by, a good scoring system devised from the ground-up, should be an example of an innovative solution.

Reviewer #2
I think they will be able to suggest a formula/method without being all that creative

Reviewer #3
The ability to display innovative thinking is somewhat limited by the overall structure of the assessment, because the questions funnel students through what they are to do at each

step.

Reviewer #5
I wonder if the disposition here should be logical while the disposition for the previous question should be innovative. The model should logically flow from the innovation of the scoring system.

6. For each of the five students in the table, calculate and report their score of study effectiveness using your method described earlier. Also, place these students in rank order according to their study effectiveness. A table is provided below to help you with this process.

	Al	Barbara	Carl	Deborah	Ed
Question 1					
Question 2					
Question 3					
Question 4					
Question 5					
Question 6					
Question 7					
Question 8					
Question 9					
Question 10					
Question 11					
Question 12					
Question 13					
Question 14					
Score					

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
*Cognitive Processes of Statistical Problem-Solv*ing – Transforms the raw data into an aggregate form (e.g., graphs, numerical summaries)	Reviewer #1 Reviewer #2 Reviewer #5	Reviewer #3 Reviewer #4	

Comments:

Reviewer #1
Yes, this question requires students to aggregate and organize summary data.

Reviewer #2
I don't see this at getting at different cognitive processes than in deriving the formula (on the statistics side, though does on the planning vs. executive side). So I would be more in favor of this cognitive process for the previous question.

Reviewer #3
The procedure that students are to use is prescribed within the item, so it is not really respondents that are coming up with a strategy. It is also a fairly simple procedure that is prescribed (performing computation and ordering the results), so I am not sure how much can really be learned about respondents' thinking.

Reviewer #4
Seems to ony ask for a numeric summary. What do you anticipate that students could do graphically? Will they have access to technology tools for entering data and computing numerical summaires or graphs?

Reviewer #5
Why do you funnel students into only calculating a score and not to draw graphs as well? You do the same in Q5.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Statistical Problem-Solving Behaviors – Analyzes the data Description: Fits and assesses a model to solve the problem.	Reviewer #4 Reviewer #5		Reviewer #1 Reviewer #3

Comments:

Reviewer #1
The question does not ask the students to analyze the results of their model. A separate question may be needed. For example, "Do you think your scoring model, and the resulting student ranks, provide a useful model for understanding effect study habits? Explain why or why not."

Reviewer #3
The rank ordering aspect of this seems to be mainly busywork.

Changes:

- Item was revised by taking out the ranking task in the question (Reviewer #3)
- New item created in MODEST 4 (item 8) to better measure the element of *transforms the raw data into an aggregate form* (Reviewer #4 & Reviewer #5). Thus, this element was removed from this item in MODEST 4.

7. If two students have the same score of study effectiveness, does that mean they have the same study habits, according to the content on the survey? Explain your reasoning.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Considers variation Description: Looks for sources of variability within the data by examining patterns in the variables or relationships between variables	Reviewer #1 Reviewer #4	Reviewer #3 Reviewer #5	Reviewer #2

Comments:

Reviewer #1
Yes, this is a good question for this element. It also aligns with integrating the statistical and contextual, in that students need to look beyond the scoring system to the context of the indicators to understand whether the same score can reflect different study habits. Students who consider the scoring or context separately, will fail to address this question correctly.

Reviewer #2
Again seems like formulas are being equated with models and I'm not sure I agree with that. This is more about being able to follow the steps, at least the steps they defined. So

maybe it's more about evaluating of their strategy – do they go back and fix their formula after they realize they can't calculate what they suggested. Actually the idea of evaluating their methods – wasn't that big in the MEAs and should be the focus here?

Reviewer #3
It would be valuable to have a question that prompts participants' inclination to look for sources of variability without being explicitly prompted to do so.

Reviewer #4
I like this question!

Reviewer #5
To examine patterns students should be encouraged to use graphs as well as numerical summaries. Do you hope the students will use graphs here as part of their explanation?

8. Write a report to the tutors of the study skills seminar that describes:

    - how to calculate your summary score of study effectiveness using the *Study Effectiveness Survey,* and
    - how to interpret the summary score so individual students can judge how effective their study habits are.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Statistical Problem-Solving Behaviors – Draws a conclusion Description: Interprets findings and communicates the results from the analysis. Critically judges the solution path and the final model for usefulness and meaningfulness.	Reviewer #1 Reviewer #4	Reviewer #3 Reviewer #5	Reviewer #2

Comments:

Reviewer #1
Yes, this a very good way to address this element.

Reviewer #2
I don't like the "draws a conclusion" part, but I think the statistical issue is more realizing the drawbacks to individual numerical summaries. The explanation of the reasoning might be helpful in judging their ability to critique the usefulness. I guess I would cast this one as more about variability (within person). But do you need to ask why someone might summarize each person with one number or ask them to propose something else. Or get at why students might need to know about results of other students to help them evaluate their own score. Do you mean same study habits or same effectiveness?

Reviewer #3
It may be productive to have respondents write two different documents – one for students (a simpler document) and one for tutors (a more technical document) to give respondents a chance to discuss limitations.

Reviewer #4
I really encourage you to allow them access to tech tools for creating numerical scores and displaying info in graphs if they so choose.

Reviewer #5
From the question it would not be obvious to students that you require the last part of the description as well ("critically judges … "). Suggest you include that the report has to describe limitations of the scoring system.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Integrates the statistical and contextual information	Reviewer #1 Reviewer #3 Reviewer #4 Reviewer #5	Reviewer #2	

Comments:

Reviewer #1
Yes, excellent.

Reviewer #2
Took me a while but I guess I see what you are getting at here. I do like the idea of going back and seeing whether they have answered the research question. Speaking of, I guess I was expecting the focus to be on judging the effectiveness of the study sessions in improving their

study habits. That might be an easier way to get at some good statistical issues.
I think I got off track here and my answers above might be shifted by a question….

I just think it's too confusing to call these "statistical models" I also don't think this question does enough to see if the student would make the choice to look at the aggregate rather than case-by-case. A better focus might be on giving them some results and asking them to make sense of them. Right now the question is too much overlapping what they did before. It could even be presenting a new rule and seeing if students can make sense of it and use it to make decisions, like comparing two groups of students. I don't think this question will get students to look at graphs or inferential statistics.

Changes:

- Item was revised by
  - asking to summarize the data rather than interpret the score (Reviewer #2) and
  - adding tasks of addressing the limitations of the survey and score (Reviewer #3 & Reviewer #5)
- Two elements were added to this item in MODEST 4 due to the revisions to the item:
  - *Reasons with statistical models,* and
  - *Is skeptical*.

---

9. Do you think it would be useful to provide students with a summary graph of their results on the survey? If so, what type of graph would you present and how would this be useful to the students? If not, why not?

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Cognitive Processes of Statistical Problem-Solving – Reasons with statistical models Description: Reasons with data from an aggregate-based approach rather than from an individual-based approach. Reasoning can include graphs, numerical summaries, and inferential statistics	Reviewer #1 Reviewer #3 Reviewer #5	Reviewer #2 Reviewer #4	

Comments:

Reviewer #1
Yes, I believe this question will produce some interesting responses, given the open-ended nature of scoring and the many different approaches students could take with plotting the data.

Reviewer #2
I worry the questions gives a bit too much away. I think if you gave them numbers and saw what they did with them you would learn more. Or give a weak graph and have them critique. I don't think you are really getting at the aggragate vs. individual here.

Reviewer #4
I think this is good AND give them access to tools to create them.

Changes:

- Item was removed and ideas were integrated into item 8 in MODEST 4 (Reviewer #2).

10. What concerns or reservations would you have about using the *Study Effectiveness Survey* or your score of study effectiveness within a classroom? Explain.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Individual Dispositions – Is skeptical (e.g., is this conclusion justified?)	Reviewer #1 Reviewer #5	Reviewer #2 Reviewer #3 Reviewer #4	

Comments:

Reviewer #1
Yes, this question does relate to sceptisim. I would hope my students would raise the issue of validity and reliability. For example, they would want to correlate it with marks and compare it to students' actual study habits (somehow). This question could be fleshed out

further as it also aligns with “Drawing conclusions” and “Integrating the statistical and contextual”. Having students critically evaluate results is one of the most common methods for assessing statistical thinking.

Reviewer #2
Does get at ability to critique though not really about a conclusion (though it’s only an eg, maybe include a second eg). Was it really a score of student effectiveness within a classroom or the individual scores for students? I think there are lots of things they could say about the survey, limit them to 1 or 2 key points? Maybe split this into two questions as well.

Reviewer #3
This item is really prompting respondents to be skeptical. I am not sure how valuable that is. It is more important that they have the inclination to be skeptical without being told to be skeptical.

Reviewer #4
Not sure about the context “within a classroom”. It seems that the survey is intended to be a self-assessment of students who go through the tutoring services.

Changes:

- Item was removed and integrated into item 9 in MODEST 4 (Reviewer #1 & Reviewer #2).

---

11. Are there other student or classroom characteristics you would consider examining to help refine your survey of effective study habits? Explain your reasoning.

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
General Problem-Solving Characteristics – Seeks alternative explanations Description: Seeks alternative explanations to help explain some response.	Reviewer #4	Reviewer #1 Reviewer #3 Reviewer #5	Reviewer #2

Comments:

Reviewer #1
Yes, to some degree. However, I think this element is better covered by Question 10. Question 11 could relate to "Consideration of variation", and "Planning".

Reviewer #2
Are you going back to what they said in Q1 or having them do more critique of the one you give them? I guess I think this question have been covered enough. I guess I don't see this as dealing with alternative explanations for a result.

Reviewer #3
This is a good item, but it does not seem to fit the description. The description talks about analyzing individuals' responses to items, but no response is really given to analyze here.

Reviewer #4
Just think about the phrase "student or classroom characteristic" and how that may be interpreted.

Reviewer #5
A problem with measuring this element is that you are prompting them with the question to consider alternative explanations. In research this is always a very hard problem to resolve.

Changes:

- Item was revised to better elicit the element of *seeking alternative explanations* (Reviewer #2, Reviewer #4, & Reviewer #5).

---

12. Did you wonder what student population the *Study Effectiveness Survey* would apply to? If so, what were your thoughts? If not, why not?

How much do you agree or disagree with the following statement?
The assessment item measures the specified statistical thinking element.

Statistical Thinking Element	Agree	Agree, but with reservations	Disagree
Individual Dispositions– Is curious and aware Description: Considers new ideas or information	Reviewer #4 Reviewer #5	Reviewer #2 Reviewer #3	Reviewer #1

that conflict with own
knowledge or assumption.

Comments:

Reviewer #1
This is a difficult element to assess. In a way, this disposition is reflected across all questions. I'm not sure how I would respond to this question. I think I would come back to defining the population implied by the research context. I don't know how I would elaborate.

Reviewer #2
Again, I like where you are going, but it's still being a bit leading. Maybe go back to the given instrument and ask one question they would want to ask the person who wrote the instrument? Or give a conclusion and ask them to think about whether it's justified. Or give a follow-up question and ask why it might be of interest.

Reviewer #3
My main concern here is that the item really prompts respondents to be curious. That is fine in an instructional activity, but for an assessment it is more valuable to know if they exhibit curiosity without being prompted to do so. I think it is important to find ways to assess curiosity and skepticism without giving it away that they should think about being curious and skeptical.

Reviewer #5
Again you are prompting to think about this. Hopefully they will answer truthfully.

Changes:

- Item was revised to better elicit the element of *is curious and aware* (Reviewer #1, Reviewer #2, Reviewer #3, & Reviewer #5).

Appendix D3: Expert Reviewer Comments regarding MODEST as an Assessment of Statistical Thinking and Subsequent Changes Made to MODEST

Part 2. Evaluation at the assessment level.

How much do you agree or disagree with the following statement?

	Agree	Agree, but with reservations	Disagree
Overall, the assessment appears to measure statistical thinking, based off of the test blueprint.	Reviewer #1	Reviewer #3 Reviewer #5 Reviewer #4	Reviewer #2

Comments:

Reviewer #1
MODEST requires students to engage in a statistical problem-solving activity, and therefore, considers the most important element of statistical thinking. If I could make suggestions for improvements, I would consider taking away the news article as its does a lot of the initial thinking for the student. This will increase variability in responses, which, from an assessment perspective, is a good outcome. You will also encounter limitations with the applicability of the problem context for use in discipline specific introductory courses, e.g. engineering, chemistry, biostatistics etc. As you know, I am an advocate for this type of assessment method, but you will need to show how this approach can be applied across contexts. As I have outlined in my feedback, statistical thinking is holistic and multidimensional. It's very difficult to map a single thinking element to a single question. One strategy might be to acknowledge this and argue that a question most closely aligns with a single element, but also shares a relationship to these other elements. You might be able to represent this as a complex network of interrelated thought patterns and mediating dispositions. At the heart of statistical thinking is an understanding of the omnipresence of variability. This is the main distinction between "statistical" thinking and other types of thinking, e.g. critical thinking. MODEST addresses this element to a degree, but I wonder if more could be included. I'm certain there must be a reason, but I wonder why sampling variability and statistical inference do not appear to be directly addressed in the task. Students could raise these concepts in some responses, but there appears to be no deliberate question that prompts students to consider this important issue. I wonder if the scenario could include questions that have students consider ways in which they could determine the validity of their survey. For example, "Propose a way to test whether your survey and scoring system does what it is designed to do." This will have the students think about validity and evaluation. They will have to propose a new investigation, which sees the problem come full circle, and

perpetuates into new problems. An evaluation study would also present the opportunity to have students discuss sampling and statistical inference and reasoning with statistical models. I think the initial MODEST test is a great start. Assessing statistical thinking is hard. I know you won't be able address or solve all my suggestions, but I thought I would share my ramblings with someone who will care ☺ Keep up the great work and I am excited to see your research unfold.

Reviewer #2
I think it's close but there are some mismatches with the cognitive processes. I think measuring dispositions will be especially difficult in such an instrument. Actually it's the claims of the skills measured by a question I disagreed with more than the questions themselves…

Reviewer #3
I would suggest having fewer questions and doing less scaffolding if this is to serve as a summative assessment. As currently written, it is a nice instructional activity, but summative assessments serve a different purpose. A summative assessment needs to assess whether respondents display dispositions such as skepticism and curiosity naturally rather than prompting and scaffolding them toward the dispositions. Similar comments apply to improving this as an assessment of statistical cognition. Over-scaffolding on a summative assessment limits the ability to assess what respondents can do without assistance.

Reviewer #4
I think you should do some rewording throughout to help potential elicitation fo the notion of considerations of variation.

Reviewer #5
A general problem solving behavior that was not included was "understanding the problem". Many statisticians say they need to spend a lot of time on this before starting to devise a plan. A disposition linked to understanding the problem is curiosity. As someone who has not taken a study skills class I spent a lot of time trying to understand the problem. (Also what does "creating an effective class" mean in the online news article?) I also spent time thinking about how to define "effective" in this context and what was a "study habit" in order to determine relevant measures for the problem. Somehow your MODEST does not capture these elements of statistical thinking. The actual study effectiveness survey also confused me. Perhaps you need to put in more detail about what the students were being asked about. As students in a school I think they would be doing at least five courses (e.g., English, Maths, History, Chemistry, Physics). What course or class was the survey referring to? Or was the survey referring to all their courses? There is also an assumption that the course is divided into chapters and has reading material. It seems the students were being asked about one particular course. Therefore the survey should have a question indicating what course they were answering the survey questions

about. Or you could ask the students whether they have any questions about the survey and why they want to know those questions before they move to Part II. You could find out whether they have the statistical problem cognitive process or disposition to seek out information, ask questions, or critique information that is given to them. Your Q12 picks up part of this problem but not all, as before you can analyse data you need to understand the survey questions and the context. I think you have done some very good work in trying to assess and measure statistical thinking – a very difficult research problem to tackle.

Changes:

- Two new items on statistical inference were added to the end of the assessment (Reviewer #1)
- Problem context was clarified (Reviewer #5)

Appendix E
Rubric for MODEST

Appendix E1: Final Rubric for MODEST

1. Think about the different factors that might affect student study habits (such as student characteristics, course characteristics, environment characteristics, etc.). Create a list of questions that you think should be on the survey to help you understand student study habits. To help you create the list, you may use the article provided, your own experience, and/or other external resources. For each question, briefly explain why student responses to that question would be useful when examining student study habits.

Element [Develops a reasonable plan]: Lists factors that could reasonably be associated with student study habits AND provides a reasonable explanation for each of the factors.

Potential factors could be:

- Study Environment (e.g., distractions)
- Social Life (e.g., work, extra curricular, facebook, going out)
- Health (e.g., sleep)
- Interest in subject matter (e.g., going to office hours, extra reading)
- Organization (e.g., plan, goal, time of day to study)
- Credit load
- Amount of time spent studying
- Studying strategies (e.g., asking questions, making connections, create drawings)
- Metacognitive (reflecting on your own learning)…Understanding own learning style and where study best

Note: Both the question and the explanation about why the factor would be useful when examining student study habits must seem reasonable and be student-related (i.e., not teacher-related).

<u>Scoring for Question 1</u>

This element is scored as follows (E, P, or I):

- <u>Essentially demonstrates (E)</u> develops a reasonable plan if…
    - **ALL** of the factors listed in the response seem reasonably associated with student study habits **AND**

- **At least (around) 80%** of the explanations to each factor in the response **CLEARLY** describe why it would be useful when examining student study habits. The explanation must clearly describe how the factor relates to student study habits.

Example answers:

- Are you in a group of friends that tends to go out a lot? (it would distract the student from studying)

- Do you have a job? If yes how many hours do you work a week? Working can cause study habits to be much different due to lack of time causing more intensive studying or not enough studying at all depending on how many hours are worked and the type of job.

- Partially demonstrates (P) develops a reasonable plan if

  - **At least SOME** of the factors listed in the response seem reasonably associated with student study habits **AND/OR**
  - **Less than 80%** of the explanations to each factor in the response reasonably describe why it would be useful when examining student study habits. Recall that the explanation must clearly describe how the factor relates to student study habits.

Example answers:

- Where do you study most effectively? If people study best in classroom environments providing that service would help the student.

  Comment: Factor is good but explanation doesn't describe how the factor relates to student study habits (i.e., no explanation).

- To understand the courses: How many credits is the course? How many hours do you spend outside of the course studying? Is there group work? Is the classroom comfortable to be in? Is the professor attentive and approachable? How many students are in the class? Is there a TA? Are your questions answered adequately? / To understand the student: How many hours a week do you work? How many hours of sleep do you receive? How well are you at taking notes? How well are you at test-taking? Do you do well with group work? Have you found study methods that work for you or are you constantly trying out new things to find something that works for you?

  Comment: Several of the factors are not related to student study habits (see underlined questions) AND no explanations are provided.

- 1. What level of interest do you have for the course in question? / 2. What is your opinion of the professor that teaches the course? Do you like them or dislike them? / 3.

Where do study most often? Do you study in your dorm room, the library, a coffee shop, etc? / 4. Would you classify yourself as an active or passive learner? Do you ask yourself questions about the lecture and homework? / 5. Is the course part of your major or something you're interested in, or is it something to fulfill a requirement or something you are disinterested in? / 6. When you sit down to study, do you have everything you need, such as your textbook, paper and pen, etc? / 7. What types of note-taking strategies do you make use of? Do you take notes at all when you study? / 8. Do you quiz yourself and what you have just read or studies?

Comment: Factors are good but no explanations are provided.

- Does not demonstrate (I) develops a reasonable plan if the response does not meet the criteria for E or P. (e.g., all factors and explanations are non-student related).

Example answer:

- Well, personally I was diagnosed with ADD in second grade, so focusing has always been extremely difficult for me, whether it be in class lectures, class readings, taking exams, etc. ADD is something that is rarely taken seriously nor acknowledged in situations similar to this. Aside from that, in modern times students are faced with distractions that constantly surround them- TV, Facebook, Twitter, Instagram....their phones in general; and it does not help the issue that many homework assignments and readings are now online. In fact, it has gotten so out of hand that applications have been developed where a student may block a social network from being accessed on their computer for a desired amount of time.

2. Do you think that the effect of the study habit factors on course grade would be the same for all students in the course? Explain why the factors would or would not have the same effect on course grade for all students.

Element [Considers variation]: Recognizes the person-to-person variation in study habits and course grades.

Scoring for Question 2

This element is scored as follows (E, P, or I):

- Essentially demonstrates (E) considers variation if the response provides a **CLEAR** indication of considering variation. For example, if the response indicates…
    - The effect **would not** be the same for all students **AND**
    - Provides a **CLEAR** explanation on why the factors would not have the same effect on course grades for all students. The explanation must say something about how the factors relate to the course grade.

Example answer:

- I believe that every student studies differently, and that it is irresponsible to believe that one study method would work for every student. While one student may have poor study habits - they cram the night before in a crowded restaurant surrounded by their loudest friends - that does not necessarily mean that they will do worse on an exam than a student that has been studying for a month. / / The way in which students retain information varies from person to person, meaning that their study methods cannot be used as direct correlations to their performance on an exam. There could be many external factors effecting their performance, such as a death in the family the morning of the test, which may negate the studying that they did. Students may be studying incorrectly for their brain type - reading all night when they are actually visual learners. They may have studied hard in a way that works for some, but because it didn't work for them, they may fail.

- Partially demonstrates (P) considers variation if the response provides **SOME** indication of considering variation. For example, if the response indicates…
    - OPTION 1:
        - The effect **would not** be the same for all students **AND**

- Provides a **VAGUE** or **LACK OF** explanation on why the factors would not have the same effect on course grades for all students. The explanation might include something about how the factors vary amongst students but doesn't relate the factors to the course grade.

Example answers:

- No. Not at all. In many ways, studying is like stretching: an activity that is presumed to be healthy and prudent and yet has a tenuous connection to academic success. Bigger, more dramatic factors that would determine students' success would be things like attitudes toward academic success, cultural background, history with math courses, natural ability, resilience and work ethic, etc. A cookbook study guide is only as good as the student who chooses to follow it.

Comment: Indicates that effect would not be the same AND explanation is vague.

- No, The effect would not be the same for all students. Every student's learning habits are different.

Comment: Indicates that effect would not be the same AND no explanation is provided.

OR

- OPTION 2:
  - **A vague answer** that the effect would not be the same for all students **AND**
  - Provides **some** indication that students vary in their study habits.

Example answer:

- I think that somethings in where you study would be different but in the end looking at the habits of how studying happens comes into play for everyone and something they can be positive or negative depending on the habit.

- <u>Does not demonstrate (I)</u> considers variation if the response does not meet the criteria for E or P.

Example answers:

- The top 3 study habits that factors on course grade in my belief is if they work, how much they sleep and if they eat breakfast. Generally if you have the free time, you dont have an excuse to not be studying. If you sleep well enough and eat well enough your body is also going to be in a good state to study so that helps everybody.

- I believe that the effect of the study habit factors on course grade would be the same for all students because all the same factors should have the same outcome.

3. In order to create the survey and use it to evaluate student study habits, what other kinds of information would you need to consider? Be sure to explain your reasoning.

Element [Develops a plan]: Lists additional kinds of information that would need to be considered in order to create and use the survey [for college students].

Potential kinds of information could be:

- Additional factors not listed in question 1
- Format of survey questions
- Length of the survey
- Inferential issues (sample collected vs. population of interest)
- Survey issues (response bias, self-reporting)

Scoring for Question 3

This element is scored as follows (E, P, or I):

- Essentially demonstrates (E) develops a plan if…
    - **ALL** of the additional information listed seem reasonable when creating and using the survey [for college students] **AND**
    - **At least (around) 80%** of the explanations listed in the response **CLEARLY** describe why it would be useful to consider when creating and evaluating the [college] student study habits survey.

Example answer:

You would need to consider how many students would actually take the survey, and what sort of biases would be associated with the survey such as self-reporting bias. This could skew the data about evaluating student study habits.

- Partially demonstrates (P) develops a plan if
    - **At least SOME** of the additional information listed seem reasonable when creating and using the survey [for college students] **AND**
    - **Less than 80%** of the explanations in the response reasonably describe why it would be useful to consider when creating and evaluating the [college] student study habits survey.

Example answers:

- I think other information that would need to be considered would be: / -Student's overall GPA (shows how strong of student they are as a whole) / -Test taking practices (shows how they exam--if they have testing anxiety, etc) / -Other factors that they feel effect their study habits (such as personal life, work life, time they have to commit).

Comment: Information is good AND several of the explanations aren't clear on why the information would be useful.

- How many hours the student studies each week.

Comment: Information is good AND no explanation is provided.

- Does not demonstrate (I) develops a plan if the response does not meet the criteria for E or P.

Example answer:

- I don't know. Is this more of Question 1?

4. When developing an overall score of study effectiveness, you will need to decide **how** each question on the *Study Effectiveness Survey* will or will not contribute to the overall score. Use the list of questions in the table below to describe **how** each question will or will not contribute to the overall score of study effectiveness. Be sure to give enough detail so that someone else could easily understand your thought processes that went into creating the overall score.

Element [Produces a conceptual model]: Describes the degree of contribution (e.g., greatly contribute, mildly contribute, not contribute) for each question on the survey AND hypothesizes about how the question topic relates to study effectiveness

Scoring for Question 4

This element is scored as follows (E, P, or I):

- Essentially demonstrates (E) produces a conceptual model if…
    - Some sort of degree of contribution **IS** articulated (e.g., greatly contribute, mildly contribute, not contribute) for each question on the survey **AND**
    - Makes a hypothesis for **ALL** of the questions about how the question topic relates to study effectiveness **AND** their hypothesis matches their degree of contribution.

Example answer [Actual Student's Response]:

	Question Content	
Question 1	**Difficulty of Material**	Huge contribution. Easier subjects count less towards overall score [of study effectiveness].
Question 2	**Prior Knowledge**	Huge contribution. Prior knowledge = higher overall score.
Question 3	**Current Grade**	Huge contribution. Higher grade = better overall score.
Question 4	**Grade as a Reflection of Learning**	Minimal contribution. Grade doesn't determine how well someone understood the material.
Question 5	**Time Spent Studying**	Minimal contribution. Effectiveness of study > time spent.
Question 6	**Distracted when Studying**	Huge contribution. Distracted studying = lower overall score.
Question 7	**Skipping Parts**	Huge contribution. Skipping parts = lower overall score.
Question 8	**Read More than Once**	Minimal contribution.

		One time could be enough for someone and not enough for the other.
Question 9	**Skim for Big Picture**	Minimal contribution. Depends on what is being learned.
Question 10	**Notetaking when Reading**	Slight contribution. Some need to and others don't.
Question 11	**Read and Not Comprehend**	Huge contribution. Will result in lower overall score.
Question 12	**Synthesize the Readings**	Huge contribution. Increase knowledge of material = higher overall score.
Question 13	**Discuss with Others**	Huge contribution. Increase knowledge of material = higher overall score.
Question 14	**Make Connections**	Huge contribution. Increase knowledge of material = higher overall score.

- Partially demonstrates (P) produces a conceptual model if…
  - OPTION 1:
    - Some sort of degree of contribution **IS** articulated (e.g., greatly contribute, mildly contribute, not contribute) for each question on the survey **AND**
    - Makes a hypothesis for **SOME** of the questions about how the question topic relates to study effectiveness **OR LACKS** a hypothesis about how the question topic relates to study effectiveness **OR** their hypotheses match only **SOME** of their degrees of contribution.

Example answers for (P) OPTION 1:

Q1:Difficulty "This will be important because some students may easily give up if material is harder"
Q2:Prior "not very essential. "
Q3:Grade "This could effect a students study habits both ways because it could either force them to study hard or make them slack off or give up"
Q4:Learning "If anything this would contribute positively if one received a good grade for studying and learning"
Q5:Time "This may not matter as some students study faster or slower than others and each are equally efficient and successful"
Q6:Distracted "Contributes greatly. distraction could greatly affect ones study time."
Q7:Skip "could be very important, depending on the material"
Q8:Read "Would vary greatly from student to student as some may need to read more than once in order to fully understand."
Q9:Skim "negative contribution"

Q10:Notetaking "very effective contribution for the majority of students"
Q11:NoComprehend "very negative study habit "
Q12:Synthesize "Im not sure what this means"
Q13:Discuss "this can be useful, however this completely depends on the study preferences of the student"
Q14:Connections "This would follow the same as question 13"

Comment: Degree of contribution is articulated AND only provides some hypotheses for each question topic.

Q1:Difficulty "It must be considered, because it brings in the human factor. However it is hard to actually tell how it affects things in numbers. "
Q2:Prior "Very important, will be looked at in conjunction with difficulty. "
Q3:Grade "Will be looked at in conjunction with question 5. however, not very important. "
Q4:Learning "This question will not factor into overall score."
Q5:Time "Will be important"
Q6:Distracted "Will be important"
Q7:Skip "Will be important"
Q8:Read "Will not be a huge factor"
Q9:Skim "will not be a huge factor"
Q10:Notetaking "Will not be a huge factor"
Q11:NoComprehend "will be weighed heavily"
Q12:Synthesize "will be weighed heavily"
Q13:Discuss "will be weighed heavily"
Q14:Connections "will be weighed heavily"

Comment: Degree of contribution is articulated AND LACKS hypotheses for each question topic.

OR

- OPTION 2:
  - Some sort of degree of contribution **IS** articulated (e.g., greatly contribute, mildly contribute, not contribute) for each question on the survey **AND**
  - Makes a hypothesis for **ALL** of the questions about how the question topic relates to study effectiveness **BUT** their hypotheses match only **SOME** of their degrees of contribution.

Example answer for (P) OPTION 2:

Q1:Difficulty "15% of the score would be based on the rated difficulty of the material because if the student has no trouble with the class then their studying will be clearly different from a student who is struggling through a course."

Q2:Prior "10% of the score would be based on the amount the student already knew about the course because if the student has no trouble with the class due to prior knowledge then their studying will be clearly different from a student who is struggling through a course."
Q3:Grade "4% of the score would be based on this element because a grade is subject to rapid change and fluctuation."
Q4:Learning "10% if the student feels as though the grade does not represent their efforts, then the studying might still be effective, but the testing environment or some other factor could be the issue for the student."
Q5:Time "5% because the amount of time spent studying could have a different impact on every student. "
Q6:Distracted "6% This is an aspect that will effect the overall effectiveness of the students studying regardless of their comfort in the class"
Q7:Skip "5% The student could be skipping parts of the material because they are comfortable with their level of understanding or they could be skipping the material because it is simply too hard for them. Because this area is so diferent for every student I would keep the percentage low."
Q8:Read "5% Some students do not learn mainly from reading so this is not representative of their overall studying habits for effectiveness"
Q9:Skim "5%This is a helpful study habit for all students, even if they feel comfortable. It also shows willingness to put in effort outside of the class."
Q10:Notetaking "4% Some students learn better by simply reading the material instead of breaking it up while they read by taking notes. I would probably ask a different set of questions to learn more about their notes and reading habits."
Q11:NoComprehend "6% If the student feels as if they are not understanding what they are reading then that is a pretty good reason to increase studying or modify reading and studying habits."
Q12:Synthesize "5%-I would want to know how the student is doing this and how often"
Q13:Discuss "10% This is a good indication of students willingness to learn the material and ask questions and possibly teach other students which to me shows effective study habits"
Q14:Connections "10% If the student is taking the time to make connections during lecture or reading, assignments, or studying then they are being effective in their approach to studying"

OR

- OPTION 3:
  - Some sort of degree of contribution is **NOT** articulated (e.g., greatly contribute, mildly contribute, not contribute) **AND**
  - Makes a hypothesis for at least **SOME** of the questions about how the question topic relates to study effectiveness.

Example answer for (P) OPTION 3:

Q1:Difficulty "Helps to see how often a student will study depending on the difficulty they find the course to be."
Q2:Prior "If they do not have prior knowledge they will probably need more study time."
Q3:Grade "To see the correlation of their current grade and their study habits."
Q4:Learning "To have the student look at their grade and analyze if it makes sense to how much work they have put into the course."
Q5:Time "To check in and see how much time they actually spend studying."
Q6:Distracted "To notice how often they have electronics and other distractions out while studying."
Q7:Skip "To become aware of habits they may not realize they have."
Q8:Read "To realize that going over information more than once is essential when studying."
Q9:Skim "To skim the material before the lecture in order to have a better grasp on the information."
Q10:Notetaking "To make a concious effort to write things down so that it makes more sense when studying."
Q11:NoComprehend "To notice how often they are reading but not saturating the information."
Q12:Synthesize "To simplify the information so that it makes sense to them."
Q13:Discuss "To talk it over with others in order to make more sense of the readings and get other opinions."
Q14:Connections "To notice how each chapter or subject studied has connections through out."

- Does not demonstrate (I) produces a conceptual model if the response does not meet the criteria for E or P (e.g., appears to fill out the survey for themselves).

Example answer for (I):

Q1:Difficulty "\"1-5\""
Q2:Prior "\"1-5\""
Q3:Grade "\"1-5\""
Q4:Learning "\"1-5\""
Q5:Time "\"1-5\""
Q6:Distracted "\"5-1\""
Q7:Skip "\"5-1\""
Q8:Read "\"1-5\""
Q9:Skim "\"1-5\""
Q10:Notetaking "\"1-5\""
Q11:NoComprehend "\"5-1\""
Q12:Synthesize "\"1-5\""
Q13:Discuss "\"1-5\""
Q14:Connections "\"1-5\""

5. Using the answers you gave in question 4, describe how to compute an overall score of study effectiveness for a student.

Element #1 [Translates the conceptual model into a statistical model]: Is able to take the ideas from the conceptual model and convert them into a single numerical (statistical) score.

Element #2 [Produces a quality model]: Is able to adequately describe how the numerical score will be computed and takes into account reverse coding for some of the survey questions.

*Note: The conceptual OR statistical model could be **described in either Q4 OR further described in Q6.** If the conceptual or statistical model is described in Q6, use Q6 to assess these elements AND make a note of it for the student.*

Scoring for Question 5

Element #1 is scored as follows (E, P, or I):

- Essentially demonstrates (E) translates the conceptual model into a statistical model if the description of how to compute a numerical score takes into account **ALL** of their ideas in the conceptual model.
    - For example, does the numerical description match the thoughts articulated in the conceptual model?

- Partially demonstrates (P) translates the conceptual model into a statistical model if…
    - OPTION 1:
        - The description of how to compute a numerical score **SOMEWHAT** takes into account their ideas in the conceptual model.
            - For example, identified questions in Q4 that would negatively contribute to study effectiveness and doesn't take that into account in their numerical score.

OR

- OPTION 2:

- **DOESN'T DESCRIBE** how to compute a numerical score **BUT** appears to try to use their conceptual model.

- Does not demonstrate (I) translates the conceptual model into a statistical model if the response does not meet the criteria for E or P.

Element #2 is scored as follows (E, P, or I):

- Essentially demonstrates (E) produces a quality model if…
    - The computation of the numerical score **IS** adequately described (e.g., average, sum) **AND**
    - Reverse coding of **SOME** of the questions on the survey are taken into account (Survey Questions 2, 6, 7, 11).

- Partially demonstrates (P) produces a quality model if…
    - The computation of the numerical score **IS** adequately described (e.g., average, sum) **AND**
    - **NO** reverse coding of any of the questions on the survey are taken into account (Survey Questions 2, 6, 7, 11).

- Does not demonstrate (I) produces a quality model if the response does not meet the criteria for E or P.

6. Use your method described in questions 4 and 5 to compute an overall score of study effectiveness for Al on the handout. Explain how you calculate the overall score for Al.

7. Fill Al's result in the table below. Then, repeat the process of using your method to calculate and report an overall score of study effectiveness for the four remaining students on the handout. Use the table below to help you with this process.

	Question Content	Al	Barbara	Carl	Deborah	Ed
Question 1	**Difficulty of Material**					
Question 2	**Prior Knowledge**					
Question 3	**Current Grade**					
Question 4	**Grade as a Reflection of Learning**					
Question 5	**Time Spent Studying**					
Question 6	**Distracted when Studying**					
Question 7	**Skipping Parts**					
Question 8	**Read More than Once**					
Question 9	**Skim for Big Picture**					
Question 10	**Notetaking when Reading**					
Question 11	**Read and Not Comprehend**					
Question 12	**Synthesize the Readings**					
Question 13	**Discuss with Others**					
Question 14	**Make Connections**					
Score						

NOTE: Q6 & Q7 will be assessed together.

Element [Analyzes data]: Applies the statistical model to the data to compute a result.

Tips for how to assess Q6 & Q7:

- Review Q5 for a particular ID (and sometimes Q4 if model description isn't clear enough or Q6 if the model description was in that item)
- Look over Q6 to get idea if followed method described in Q5
- Look over Q7 to see if applied method (correctly) across several students

Note #1: Q6 was included in the assessment to try to understand a student's analysis of the data if can't follow their logic in Q7. Primarily use Q7 to assess the element of "analyzes data."

Note #2: Focus on the total score computed and not the values placed in the table.

Scoring for Question 6 & 7

This element is scored as follows (E, P, or I):

- Essentially demonstrates (E) analyzes data if the response **accurately applies** their statistical model (from Q5 or Q6) to the data to compute a result (i.e., a total score) **for at least 3 of the 5 student data.**

- Partially demonstrates (P) analyzes data if the response **accurately applies** their statistical model (from Q5 or Q6) to the data to compute a result (i.e., a total score) **for at least 2 of the 5 student data.**

- Does not demonstrate (I) analyzes data if the response does not meet the criteria for E or P. (e.g., adds up all of the survey responses from the student data but doesn't describe statistical model in Q5 or Q6).

8. If two students have the same score of study effectiveness, does that mean they have the same study habits, according to the content on the survey? Explain your reasoning.

Element [Considers variation]: Recognizes the person-to-person variation in the study habit factors (according to the survey) in relation to the score of study effectiveness.

Scoring for Question 8

This element is scored as follows (E, P, I):

- Essentially demonstrates (E) considers variation if the response provides a **CLEAR** indication of considering variation. For example, if the response indicates…
    - The study habits **may not be the same** for the students who receive the same score of study effectiveness **AND**
    - Provides a **clear** explanation on why similar study habits scores don't mean same study habits.

Example answer:

- It means that have some similarities, but there are always differences in habits. We know they have a similar time spent studying, but it could be all in one sitting or a little each day. They could also have higher scores in different sections.

- Partially demonstrates (P) considers variation if the response provides **SOME** indication of considering variation. For example, if the response indicates…
    - OPTION 1:
        - The study habits **may not be the same** for the students who receive the same score of study effectiveness **AND**
        - Provides a **VAGUE OR LACK OF** explanation on why similar study habits scores don't mean same study habits.

Example answers:

- It does not mean that they have the same study habits at all. Just that the amount of improvement they both need is similar, even if its in different areas.

Comment: Indicates that study habits may not be the same AND provides a vague explanation.

- No. It means they have the same effectiveness.

Comment: Indicates that study habits may not be the same AND no explanation is provided.

OR

- OPTION 2:
  - **A vague answer** that the study habits may not be the same for the students who receive the same score of study effectiveness **AND**
  - Provides **some** indication that students vary in their study habits.

Example answer:

- Not necessarily...since each question talks about a different means of studying you can't assume that the same score means that the two students score the same.

- Does not demonstrate (I) considering variation if the response does not meet the criteria for E or P.

Example answer:

- Maybe. Because they have similar results.

9. If the instructor used her survey to collect data from her whole class, describe the statistical measures or methods that you would use to provide a summary of the effectiveness of their study habits.

Element [Appropriately reasons with statistical models]: Lists names of specific descriptive statistical measures or methods (e.g., mean, sd, median, correlation, histogram) that would be useful in summarizing the study habit scores.

Scoring for Question 9

This element is scored as follows (E, P, or I):

- Essentially demonstrates (E) appropriately reasons with statistical models if the response **names specific descriptive statistical measures or methods** (e.g., mean, sd, median, correlation, histogram) would be useful in summarizing the study habit scores (at the class level, not the student level) with **enough detail** that you can picture what the report would look like.

Example answers:

- I would use key statistical concepts such as the mean, mode, and median, in order to easily show the class what the average scores were looking like, as well as the most common scores and the middle scores. The range would also be helpful too, just to show if there were any outliers are things that would skew the data.

- Show class average score, with highest and lowest included as frames of reference. Break down individual questions into pie charts showing the % of the class that was in which range.

- Partially demonstrates (P) appropriately reasons with statistical models if the response…
    - Option 1:
        - **Lists descriptive statistical measures or methods** (e.g., graphs, tables, bar graphs) that would be useful in summarizing the study habit scores with **vague detail** that you can't adequately picture what the report would look like.

Example answers:

- I would provide a summary using table and charts based on the data collected.

- I would do a bar graph because its easy to understand and look at…

Comment: Didn't describe the variable that would be plotted on the bar graph or used in the tables and charts so the students provided vague detail.

OR

- Option 2:
  - **Lists descriptive statistical measures or methods** (e.g., graphs, tables, bar graphs) that would be useful in summarizing the study habit scores **AND** also **mentions using a confidence interval or standard error**.

Example answers:

- The statistical measures that I would use to provide a summary of the effectiveness of the whole class's study habits would be the mean in order to provide an average of the scores, and an interval created by finding the standard error, multiplying it by two to get the margin of error and taking that plus or minus the mean, in order to be able to say where the majority of study effectiveness scores lie.

- After using the survey to collect data from her whole class, I would use my measurement strategy to determine the average level of study effectiveness in her class. I could bootstrap this data and run 500 trials (something that would be impossible in her actual class) and see if the trend stays the same and then apply it to the population of her class. I could use this analysis to show the instructor how well her students are studying. This could be compared to the average grades in the class and the scores on exams to see where issues lay in the course.

- Does not demonstrate (I) developing a plan if the response does not meet the criteria for E or P. For example, only states a confidence interval technique (e.g., bootstrapping) or mentions using hypothesis testing (e.g., randomization test).

Example answers:

- The statistical measures/methods I would use to provide a summary of the effectiveness of their study habits would be that the higher the score the better study habits the student has.

- Well you could run a randomization trial with grades received and study scores as recorded by this test! That would hopefully give you a realistic P-value to work with and make inferences from. You would definitely need to refine the multipliers and method of the study effectiveness survey first, however.

- We could use my method above because it pinpoints those who are generally not doing too well or enough of what they need to do. However, it does yield results that are questionable due to confounding variables and misconceptions of my method versus the student's answers. I would say that after bootstrapping these results, we can draw on a generalization due to random sampling. If we yield these results, it would work that the plot shows parameters for effectiveness of studying.

10. Write a brief report (~1-2 paragraphs) to the instructor that addresses:

- how the overall score of study effectiveness was calculated using the *Study Effectiveness Survey*,
- how to interpret an overall score of study effectiveness,
- a summary of the overall scores for the five students,
- the potential limitations of the *Study Effectiveness Survey*, and
- how well you think the overall score measures study effectiveness.

Element #1 [Draws a conclusion]: Provides a reasonable description of how to interpret the overall score of study effectiveness.

Element #2 [Appropriately reasons with statistical models]: Describes a summary for the five student results of overall scores of study effectiveness that would be of use to the client (e.g., aggregate-based approach [numerical summaries] vs. individual case-based approach).

Element #3 [Is skeptical (critical)]: States reasonable limitations of the survey AND provides a thoughtful critique of using the overall score as a measure of study effectiveness.

Note: Needed to sometimes examine their response in Q7 to see what actually did and what they calculated for the students to see if they are making the correct conclusions.

Scoring for Question 10

Element #1 is scored as follows (E, P, or I):

- Essentially demonstrates (E) draws a conclusion if the response provides a reasonable description of how to **interpret the overall score** of study effectiveness

Example answers:

- The statistical measures/methods I would use to provide a summary of the effectiveness of their study habits would be that the higher the score the better study habits the student has.

- Well you could run a randomization trial with grades received and study scores as recorded by this test! That would hopefully give you a realistic P-value to work with and make inferences from. You would definitely need to refine the multipliers and method of the study effectiveness survey first, however.

- We could use my method above because it pinpoints those who are generally not doing too well or enough of what they need to do. However, it does yield results that are questionable due to confounding variables and misconceptions of my method versus the student's answers. I would say that after bootstrapping these results, we can draw on a generalization due to random sampling. If we yield these results, it would work that the plot shows parameters for effectiveness of studying.

- Partially demonstrates (P) draws a conclusion if the response provides a **VAGUE** description of how to **interpret the overall score** of study effectiveness.

Example answer:

- …if their score is at 70, their score is perfect and they are doing everything right. Their scores varied from 20s to 50s. Just because someone had a low score does not mean their study habits need adjusting. It just depends on what section their scored low in…

- Does not demonstrate (I) draws a conclusion if the response does not meet the criteria for E or P.

Example answer:

- Each student answered questions that summarized their own way of studying into a number. Each student should take their surgery truthfully but every student has their different opinion on what number they fall under. Most students were in the A or B range for grades. Whatever they are doing for studying is working for them. There is one student that has a D in the class and should probably change his way of studying because it isn't work that well for that specific student…

Comment: Lack of interpretation of overall score of study effectiveness.

Element #2 is scored as follows (E, P, or I):

- Essentially demonstrates (E) appropriately reasons with statistical models if the response describes a **summary for the five student results** of overall scores of study effectiveness (e.g., aggregate-based approach [numerical summaries] vs. individual case-based approach).

Example answer:

- …The overall scores for the five students were basically reflective of their grade in the course. The student's with higher scores also presented higher grades in the course…

Comment: The summary for the student results describe an analysis via correlation (grade with overall score).

- Partially demonstrates (P) appropriately reasons with statistical models if the response **DOES NOT adequately describe a summary for the five student results** of overall scores of study effectiveness (e.g., only provides minimum and maximum values).

Example answers:

- …The five students represented a whole range of possible outcomes, each having a unique route to determine their score…

- …Their scores varied from 20s to 50s….

- Does not demonstrate (I) appropriately reasons with statistical models if the response does not meet the criteria for E or P.

Example answer:

- …Most students were in the A or B range for grades. Whatever they are doing for studying is working for them. There is one student that has a D in the class and should probably change his way of studying because it isn't work that well for that specific student…

Comment: Summarizes the student's grades and not their overall score.

Element #3 is scored as follows (E, P, or I):

- Essentially demonstrates (E) <u>is critical</u> if the response…
    - States **reasonable limitations of the survey** (e.g., questions, format), not related to the score **AND**
    - Provides a **thoughtful critique of using the overall score** as a measure of study effectiveness.

Example answer:

- …The potential limitations of the survey is that not all the questions can be assumed that a 1-5 scale will properly assess effectiveness of studying. Personally, I don't thin k that this score is very effective at telling study habits because the questions are pretty subjective.

- Partially demonstrates (P) <u>is critical</u> if the response…
    - Option 1:
        - States **reasonable limitations of the survey BUT**
        - **DOES NOT** provide a **thoughtful critique of using the overall score** as a measure of study effectiveness.

Example answer:

- …The SES (Study Effectiveness Survey) is limited because it's vague and doesn't give students the chance to explain their unique perspective. I think the SES is not extensive enough to truly reveal study effectiveness.

OR

- Option 2:
    - **DOES NOT states reasonable limitations of the survey BUT**
    - Provides a **thoughtful critique of using the overall score** as a measure of study effectiveness.

Example answer:

- …I do not think this is an effective study. Two overall scores could be the same, but those two students could be struggling with two completely different sections.

- Does not demonstrate (I) is critical if the response does not meet the criteria for E or P (e.g., lacks a critique of the survey and the overall score).

Example answers:

- I don't know.

- …The overall score of study effectiveness was calculated by using their score on the survey as an effective rating for the amount of hours spent to receive their overall grade. It can best be interpreted by a table. The overall scores show that Deborah has the most reliable score to show how much time spent studying verses overall grade.

11. What suggestions do you have to help the instructor revise the *Study Effectiveness Survey* or the overall score of study effectiveness*?*

Element [Is critical]: Is able to effectively analyze and evaluate the usefulness and meaningfulness of the definition of the problem (as defined by the questions on the survey) or the model (overall score of study effectiveness) they proposed.

Potential critiques could be:

- Survey-related (e.g., adding more questions, editing the questions, modifying the survey format, putting items all on same scale)
- Score-related (e.g., modifying score, not creating a single score)
- Construct-related (e.g., survey not measure study effectiveness, not a measureable construct, "validation" of survey with outside factors)

Scoring for Question 11

This element is scored as follows (E, P, or I):

- Essentially demonstrates (E) is critical if the response provides a **CLEAR** description of suggestions for how the survey and/or score can be revised.

Example answers:

- In order to revise the Study Effectiveness Survey, I would include more questions. If a student who is failing the class is still receiving a high score it may be because we are not asking the questions that reflect how they study. With more questions we may be able to see the areas in which they are lacking more clearly, which would make the correlation between their study habits and grade make more sense.

- I suggest to toss out this survey and look at concrete info, like GPA, course load, amount of hours studied per week, and measures that are facts, not opinions.

- Partially demonstrates (P) is critical if the response provides a **VAGUE** description of suggestions for how the survey and/or score can be revised.
    - *Note: Vague means that you are not fully able to follow their reasoning to implement their suggestions.*

Example answer:

- Allow students to prove what they've been learning through studying or look at individual test scores after studying for a test.

- Does not demonstrate (I) is critical if the response does not meet the criteria for E or P (e.g., suggests random sampling, random assignment, or larger sample).

Example answer:

You can't measure study effectiveness because every student studies at a different pace and with different habits, it's very arbitrary.

12. As a result of reading the article and working through the first 11 questions, was there anything that you wondered about regarding the evaluation of student study habits? If so, what did you wonder about? If not, why not?

Element [is curious]: Shows an interest of looking beyond the surface of the problem scenario and ponders aspects related to statistics.

Example answers of curiosity:

- Validity evidence for the construct of interest (e.g., question-level, survey-level)
- Evaluation of the construct of interest (e.g., with a score, on a survey)
- Population of interest
- Creation of the survey questions
- Survey-related aspects (e.g., administration, response bias)

Scoring for Question 12

This element is scored as follows (E, P, or I):

- Essentially demonstrates (E) is curious if the response provides a **CLEAR** description of what was pondered about during the process of evaluating students study habits, as related to statistical curiosity.

Example answers:

- I wonder if it is possible to determine this evaluation. If students share habits but have different skill sets how is that accounted for?

- I wondered why they would choose to administer this in the middle of the course rather towards the beginning. It seems to me that giving the students the survey closer to the beginning of the year would be most effective so that there would be more time to help them improve the areas they lacked in.

- Partially demonstrates (P) is curious if the response provides a **VAGUE** description of what was pondered about during the process of evaluating students study habits, as related to statistical curiosity (e.g., you have to infer from their response what statistical aspect they were curious about).

Example answer:

Yes, I wondered a lot about which questions to keep or toss. Initially all of the questions seemed important and I struggled with trying to keep them or delete them.

> Comment: This is vague because don't know if this is about their struggles completing the task or about which questions help to best describe the construct of interest.

- Does not demonstrate (I) is curious if the response does not meet the criteria for E or P. (e.g., states "is not curious", doesn't directly answer the question, provides a critique rather than ponders about something, is curious about their own answers to questions on this assessment).

Example answers:

- I wondered how I would score as I was doing this assignment.

- I just wondered what on earth I was supposed to write in the boxes on question 4 and what I was supposed to do with the five students scores.

Part IV: Extensions

The math department at the instructor's university heard that you helped in developing and using a measure of study effectiveness. They decide to hire you to help them investigate and evaluate their math students study habits.

13. The math department is interested to see if there is a difference in the study habits between students in the two math course formats. They have the following question:

    > Is there a difference in the student study habits between students who enroll in face-to-face mathematics courses and those who enroll in online mathematics courses?

    Using what you've learned in your statistics course(s), provide an brief outline of how the instructor should go about answering this question.

Element #1 [develops a reasonable plan]: Describes a plan for how to analyze the data to answer the research question posed by the math department.

Element #2 [appropriately reasons with statistical models]: Describes an appropriate statistical inferential method (e.g., confidence intervals, hypothesis testing, simulation approach) to answer the research question posed by the math department.

Element #3 [recognizes the need for data]: Indicates collecting data about student study habits to answer the question posed by the math department.

Potential data could be:

- Responses from students to questions (e.g., on a survey)
- Characteristics from aspects in a course (e.g., course grade, course attendance)

<u>Scoring for Question 13</u>

Element #1 is scored as follows (E, P, or I):

- Essentially demonstrates (E) <u>develops a reasonable plan</u> if the response describes a **reasonable plan for how to analyze the data** to answer the research question. For example,
    - Response indicates comparing the groups given the type of data they propose to collect (e.g., scores on a survey).

Example answers:

- They should construct a survey or study that is exactly the same for both groups, and compare the scores and results to see if there is a difference between the two…

- Look at the effectiveness course and compare the two groups. Compare the course grade.

- Partially demonstrates (P) develops a reasonable plan if the response describes an **incomplete plan for how to analyze the data** to answer the research question. For example,
    - Response indicates analyzing the results but doesn't indicate how the results will be compared.

Example answer:

- Ask the same 14 questions (modified for differences in the classes), then have the students self report the data, weight the questions differently according to which study habit is effective, make a scale, and analyze the results.

- Does not demonstrate (I) develops a reasonable plan if the response does not meet the criteria for E or P.

Example answers:

- Again, I don't really know. But I would present both groups with the same survey.

- Look at grades of those who took a class face-to-face and those who took the class in person. Also look whether or not online student utilized office hours even though their class was online. Provide a survey to see different study habits of the students, maybe those who took the class online don't necessarily do as much practice by themselves or didn't ask questions when they had them.

is scored as follows (E, P, or I):

- Essentially demonstrates (E) appropriately reasons with statistical models if the response **correctly describes an appropriate statistical inferential method** (e.g., confidence intervals, hypothesis testing, simulation approach) to answer the research question posed by the math department.

    - Appropriate statistical inferential methods include:
        - CI for a difference in means (e.g., traditional frequentist, bootstrap approach)
        - Hypothesis testing for a difference in means (e.g., two-sample t-test)
        - Randomization test for a difference in means

Example answer:

- I think math department should try to recruit as many participants as possible, and use an unbiased and balanced survey. Using t-test to distinguish whether the different is statistically significant or not. Then they can know the result.

- Partially demonstrates (P) appropriately reasons with statistical models if the response **incorrectly describes an appropriate statistical inferential method** (e.g., confidence intervals, hypothesis testing, simulation approach) but contains enough information that you can tell they were on the right track.

- Does not demonstrate (I) appropriately reasons with statistical models if the response does not meet the criteria for E or P.

Example answer:

- They should construct a survey or study that is exactly the same for both groups, and compare the scores and results to see if there is a difference between the two…

Element #3 is scored as follows (E, P, or I):

- Essentially demonstrates (E) recognizes the need for data if the response **CLEARLY** indicates collecting data about student study habits (e.g., via a survey, via questions, via course characteristics).

Example answers:

- They should construct a survey or study that is exactly the same for both groups, and compare the scores and results to see if there is a difference between the two…

- Look at the effectiveness course and compare the two groups. Compare the course grade.

- Partially demonstrates (P) recognizes the need for data if the response **SOMEWHAT** indicates collecting data about student study habits.

Example answer:

- I think they should employ identical study standards (time frame, group or individual, etc…) and only change the method of how the information is studied (computer as individual or in a class setting). You can employ many of the same factors in the above chart.

- Does not demonstrate (I) recognizes the need for data if the response does not meet the criteria for E or P (e.g., no data description provided, anecdotal evidence.

Example answers:

- I have not yet learned enough information to answer this question.

- Not having that one on one explanation in front of you and taking notes is not well advised for online classes. Also, not being able to ask other student for help is also hard too.

Appendix E2: Changes in the elements of statistical thinking and rubric description of the elements for each item

Item	Original Element(s)	Modification(s) to the Elements	Rubric Description of Element
1	Element 1: *Develops a plan for collection or analysis of the data.*	Element 2 was incorporated into Element 1: *Develops a [reasonable] plan*.	Lists factors that could reasonably be associated with college student study habits AND provide a reasonable explanation for each of the factors.
	Element 2: *Is logical*	Element 2 was dropped from the item.	-
2	*Considers variation*	-	Recognizes the person-to-person variation in study habits and course grades.
3	*Develops a plan for collection or analysis of the data.*	-	Lists additional kinds of information that would need to be considered in order to create and use the survey for college students.
4	Element 1: *Creates model*	Element 1 was modified to have a sub-element: *Produces a conceptual model*.	Describes the degree of contribution (e.g., greatly contribute, mildly contribute, not contribute) for each question on the survey AND hypothesizes about how the question topic relates to study effectiveness
	Element 2: *Is innovative*	Element 2 was dropped from the item.	-
5	*Creates a model*	Element was modified to have two sub-elements: • Element #1: *Translates the conceptual model into a statistical*	Element #1: Is able to take the ideas from the conceptual model and convert them into a single numerical (statistical) score.

		model, and • Element #2: *Produces a quality model*.	Element #2: Is able to adequately describe how the numerical score will be computed AND takes into account reverse coding for some of the survey questions.
6	*Analyzes the data*	Element was dropped from the item.	-
7	*Analyzes the data*	-	Applies their statistical model to the data to compute a result.
8	*Considers variation*	-	Recognizes the person-to-person variation in the study habit factors (according to the survey) in relation to the score of study effectiveness.
9	Element 1: *Reasons with statistical models*	-	Lists names of specific descriptive statistical measures or methods (e.g., mean, sd, median, correlation, histogram) that would be useful in summarizing the study habit scores.
	Element 2: *Applies previous knowledge or adapts a previous problem to fit a new problem*	Element 2 was dropped from the item.	-
10	Element 1: *Draws a conclusion*	Only the interpretation aspect of the draws the conclusion description was assessed.	Provides a reasonable description of how to interpret the overall score of study effectiveness.
	Element 2: *Integrates the statistical and contextual information*	Element 2 was dropped from the item.	-
	Element 3: *Reasons with statistical models*	-	Describes a summary for the five student results of overall scores

			of study effectiveness that would be of use to the client (e.g., aggregate-based approach [numerical summaries] vs. individual case-based approach).
	Element 4: *Is skeptical/critical*	-	States reasonable limitations of the survey AND provides a thoughtful critique of using the overall score as a measure of study effectiveness.
11	*Seeks alternative explanations*	Element was dropped from the item and replaced with *is skeptical/critical.*	Is able to effectively analyze and evaluate the usefulness and meaningfulness of the definition of the problem (as defined by the questions on the survey) OR the model (overall score of study effectiveness) they proposed.
12	*Is curious*	-	Shows an interest of looking beyond the surface of the problem scenario and ponders aspects related to statistics.
13	Element 1: *Develops a plan for collection or analysis of the data.*	-	Describes a plan for how to analyze the data to answer the research question posed by the math department.
	Element 2: *Applies previous knowledge or adapts a previous problem to fit a new problem*	Element 2 was dropped from the item.	-
	Element 3: *Reasons with statistical models*	-	Describes an appropriate statistical

		inferential method (e.g., confidence intervals, hypothesis testing, simulation approach) to answer the research question posed by the math department.
Element 4: *Recognizes the need for data*	-	Indicates collecting data about student study habits to answer the question posed by the math department.

Appendix F

Materials for Student Participants

Appendix F1: Script for Recruiting Senior Statistics Students

Hello students,

I am here to invite you to participate in a research project that is developing an assessment of statistical thinking, called Modeling To Elicit Statistical Thinking (MODEST). Statistical thinking usually means thinking like a statistician.

The development of this instrument is part of my doctoral dissertation in Statistics Education at the University of Minnesota

I am inviting you to participate in an interview that is designed to help me see how you reason and interpret statistical questions. During this interview, you will be asked to talk aloud as you solve the problems. Your responses will help me understand how students are thinking statistically on the assessment and will help me improve the assessment.

The problems may not look like anything you have done before and the problem doesn't have one correct answer. You do not have to review anything prior to the interview.

As an incentive to participate in this study, you will receive a $20 Amazon.com gift card.

I am planning to conduct the interviews from October 27th to November 7th. If you are interested in participating, please write down your name and email address in a timeslot that works for you on the sign-up sheet. I will then send you a reminder email of the date and time of your interview.

Thank you.

Appendix F2: Script for Cognitive Interviews

Read to participant:

Thanks for meeting with me. Let me tell you a little more about what you'll be doing today.

1. I am piloting a new statistics exam with the help of students, such as yourself.
2. I'll give you the exam tasks and questions and you answer them, just like a regular exam.
3. However, unlike regular exams that are done in silence, my goal here is to get a better idea of how the questions are working. So I'd like you to think aloud as you answer the questions and solve the task—just tell me everything you are thinking about as you go about answering them.
4. Please read the exam scenario, instructions, and questions aloud while you are taking the exam.
5. Please keep in mind that I really want to hear all of your opinions and reactions. Do not hesitate to speak up whenever something seems unclear or is hard to answer.
6. Sometimes I will remind you to think aloud as you are working on a task or answering a question.
7. We'll do this for about an hour and a half.
8. Please take the time to look over the consent form and sign it at the bottom.
9. Do you have any questions before we start?
10. Now let’s get started.

Think-Aloud Practice:

- Let's begin with a couple of practice questions. Remember to try to think aloud as you answer.
- Practice question 1: How many windows are there in the house or apartment where you live?
- [Probe as necessary]: How did you come up with that answer?
- Practice question 2: How difficult was it for you to get here to do the interview today: very difficult, somewhat difficult, a little difficult, or not at all difficult?
- [Probe as necessary]: Tell me more about that. Why did you say [ANSWER]?
- OK, now let's start on the exam.

Think-Aloud Interview:

- The student will be provided with the copy of the assessment.
- The student will be asked to complete the assessment while thinking aloud.
- Probes will be used if the student forgets to think aloud. Probes will not be used to elicit an answer from the student. Example probes include

- "What are you thinking?"
- "Keep talking"
- If asked what something means ask "What do you think it means?"

- After the student completes the assessment, the student will be thanked and be permitted to leave. Remind of the voluntary nature of the study and that you will document it in the subjects file.

Appendix F3: Consent Form for Cognitive Interviews

This assessment is part of a research project for developing an assessment called *Modeling To Elicit Statistical Thinking (MODEST).* We ask that you read this form carefully and ask any questions you may have before agreeing to be in the study.

This study is being conducted by Laura Le, a doctoral candidate in the Department of Educational Psychology at the University of Minnesota, under the supervision of Dr. Joan Garfield and Dr. Andrew Zieffler.

Background Information:
The aim of this study is to create an assessment of statistical thinking that attempts to measure the change (i.e., pre/post) in students' statistical thinking as a result of an introductory statistics course, at the undergraduate level. Statistical thinking has generally been considered as "thinking like an expert statistician." The assessment utilizes techniques (e.g., type of problem, open-ended questions) that have been suggested as ways of assessing expert-like thinking.

Procedures:
You will participate in a 1.5-hour interview that is designed to gain an understanding of how you are using statistical thinking on the questions in the *MODEST* assessment. Each interview will be audio-taped to produce a record of your responses for later analysis.

Consent:
Excerpts of your interview may be used in research presentations or publications as an illustration of students' statistical thinking capabilities. These excerpts may be in the form of a transcription of your statements during the interview, or of audio files selected from an interview.

We are asking for your consent regarding three things.
1. To audio-tape and record the interview.
2. To include audio files of your interviews in presentations of this research.
3. To include excerpts of your statements during the interviews in research presentations and publications.

Compensation:
You will receive a $20 Amazon.com gift certificate for your participation in the 1.5-hour interview.

Risks and Benefits of Being in the Study:
As with all research, there is a chance that confidentiality could be compromised. However, we are taking precautions to minimize this risk.

The benefit of participating is the opportunity to develop a better understanding of statistics, and of your own statistical thinking capabilities.

Confidentiality:
The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify you as a participant. All research

records will be de-identified and stored on a secure server; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study:
Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with your institution. If you decide to participate, you are free to withdraw at any time without affecting those relationships. You can quit the interview and/or the audio recording at any time and if you decide to end participation during or after the interview, you can ask for the audio recording to be destroyed.

Contacts and Questions:
The researcher conducting this study is Laura Le under the advisement of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Andrew Zieffler, Ph.D. (Educational Psychology – Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Laura Le, at free0312@umn.edu. You may also contact my advisor, Dr. Joan Garfield, at jbg@umn.edu.

If you have any questions or concerns regarding the study and would like to talk to someone other than the researchers, you are encouraged to contact the Research Subjects' Advocate line, D528 Mayo, 420 Delaware Street S.E., Minneapolis, Minnesota 55455; telephone 612-625-1650.

You will be given a copy of this form to keep for your records.

Statement of Consent

I have read the above information. I have had the opportunity to ask questions and receive answers.

Please sign and return this consent form if you agree to let us use your responses in the research study described above. Please place an X next to each item below for which you do give your permission.

	I give permission to be recorded and audio-taped.
	I give permission to include audio files of my interview in presentations of this research.
	I give permission to include excerpts of my statements in research presentations and publications.

Your Name (Please PRINT): _____

Signature: _____ Date: _____

IRB Code Number: 1409P53924
Version date: Oct. 22, 2014

Appendix F4: Script for Recruiting (Pilot) CATALST Students

Hello students,

Do you want to get EC for this course, be entered into a raffle for a $50 Amazon gift card, and help out with research?

I am here to invite you to participate in a research project that is developing an assessment of statistical thinking, called Modeling To Elicit Statistical Thinking (MODEST).

The development of this instrument is part of my doctoral dissertation in Statistics Education at the University of Minnesota

You will participate by completing *MODEST* online. The assessment consists of 12 open-ended questions and will take 50 to 70 minutes to complete. Your responses will help me understand how students are thinking statistically on the assessment and will help improve the assessment.

The problem may not look like anything you have done before and the problem doesn't have one correct answer. You do not have to review anything prior to taking the assessment.

As an incentive to participate in this study, you will receive extra credit in EPsy 3264 by having your lowest homework grade replaced with full marks. In addition, you will be entered into a raffle drawing to receive a $50 Amazon.com gift card when you complete this assessment.

You will receive an email from your instructor sometime next week. So this is just a heads up letting you know this opportunity is coming your way.

Here are some handouts that will be used on the assessment. If you lose these between now and next week, there will also be a link for these handouts in the online assessment.

Thank you.

Appendix F5

Email to (Pilot) CATALST Instructors on Administering MODEST

Dear EPsy 3264 Instructor,

Thank you for allowing me to administer the *MODEST* assessment to your students.

Below is the detailed information for administering the online assessment in your statistics course.

First, send the initial email to your students on December 10th. Copy and paste the text in the **Initial Student Email** section below into the body of an email. Be sure that the emails are blind carbon copied (BCC).

Second, send a reminder email to your students on December 15th. Copy and paste the text in the **Reminder Student Email** section below into the body of an email. Again, be sure that the emails are blind carbon copied (BCC).

Third, I will send an email to you on December 18th with the names of students from your class that completed the assessment.

Thank you again!

Sincerely,

Laura Le
Doctoral Candidate
Department of Educational Psychology
Statistics Education

Initial Student Email:

TO: EPSY 3264 students
FROM: Laura Le, Doctoral Candidate, Educational Psychology

You are being invited to participate in a research project on developing an assessment called *Modeling To Elicit Statistical Thinking* (MODEST). Your responses to this assessment are important because they will help me understand how students are thinking statistically on the assessment and will help improve the assessment.

To incent you to participate in my study, you will receive extra credit in EPsy 3264 by replacing your lowest homework grade. But that's not all! You will also be entered into a

raffle drawing to receive a $50 Amazon.com gift card when you complete this assessment.

To complete the assessment, please click on the following link:

https://umn.qualtrics.com/SE/?SID=SV_6kUSmx5J63gZAgt

The assessment consists of 12 open-ended questions and will take approximately 50 to 70 minutes to complete. You can use any resources that you want when completing the assessment except other people.

The due date for completing the assessment is **December 17th by 10p**.

If you have any questions about the assessment, please email me at free0312@umn.edu.

Thank you! Have a wonderful day!

Sincerely,

Laura Le
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

Reminder Student Email:

TO: EPSY 3264 students
FROM: Laura Le, Doctoral Candidate, Educational Psychology

Last week, you received an email with an opportunity to participate in a research project on developing an assessment called *Modeling To Elicit Statistical Thinking* (MODEST). If you have already completed this assessment, thank you!!

If you have not had the chance to complete the assessment, you have until **December 17th 10p** to take it. Your responses to this assessment are important because they will help me understand how students are thinking statistically on the assessment and will help improve the assessment.

Participating in this research project has additional benefits besides contributing to my research. You will receive extra credit in EPsy 3264 by replacing your lowest homework grade. But that's not all! You will also be entered into a raffle drawing to receive a $50 Amazon.com gift card when you complete this assessment.

To complete the assessment, please click on the following link:

https://umn.qualtrics.com/SE/?SID=SV_6kUSmx5J63gZAgt

The assessment consists of 13 open-ended questions and will take approximately 50 to 70 minutes to complete. You can use any resources that you want when completing the assessment except other people.

If you have any questions about the assessment, please email me at free0312@umn.edu.

Thank you! Have a wonderful day!

Sincerely,

Laura Le
Doctoral Candidate
Department of Educational Psychology
University of Minnesota

Appendix F6

Online Consent Form for (Pilot) CATALST Students

This assessment is part of a research project for developing an assessment called *Modeling To Elicit Statistical Thinking (MODEST*). We ask that you read this form carefully and ask any questions you may have before agreeing to be in the study.

This study is being conducted by Laura Le, a doctoral candidate in the Department of Educational Psychology at the University of Minnesota, under the supervision of Dr. Joan Garfield and Dr. Andrew Zieffler.

Background Information:
The aim of this study is to create an assessment of statistical thinking that attempts to measure the change (i.e., pre/post) in students' statistical thinking as a result of an introductory statistics course, at the undergraduate level. Statistical thinking has generally been considered as "thinking like an expert statistician." The assessment utilizes techniques (e.g., type of problem, open-ended questions) that have been suggested as ways of assessing expert-like thinking.

Procedures:
You will complete an online version of the assessment. The assessment consists of 12 open-ended questions and will take 50 to 70 minutes to complete.

Compensation:
You will receive extra credit in EPsy 3264 when you complete this assessment. The extra credit will be replacing your lowest homework grade with full marks.

In addition, you will be entered into a raffle drawing to receive a $50 Amazon.com gift card when you complete this assessment.

Risks and Benefits of Being in the Study:
As with all research, there is a chance that confidentiality could be compromised. However, we are taking precautions to minimize this risk.

The benefit of participating is the opportunity to develop a better understanding of statistics, and of your own statistical thinking capabilities.

Confidentiality:
The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify you as a participant. All research records will be de-identified and stored on a secure server; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study:
Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with your institution or the course. If you decide to participate, you are free to withdraw at any time without affecting those relationships.

Contacts and Questions:
The researcher conducting this study is Laura Le under the advisement of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Andrew Zieffler, Ph.D. (Educational Psychology – Statistics Education). If you are willing to participate or have any questions you are encouraged to contact me, Laura Le, at free0312@umn.edu.You may also contact my advisor, Dr. Joan Garfield, at jbg@umn.edu.

If you have any questions or concerns regarding the study and would like to talk to someone other than the researchers, you are encouraged to contact the Research Subjects' Advocate line, D528 Mayo, 420 Delaware Street S.E., Minneapolis, Minnesota 55455; telephone 612-625-1650.

Statement of Consent:

Please click the circle below if you agree to participate in this research study.

	I have read the above information and I give permission for my responses to assessment items to be included in any analyses, reports, or research presentations made as a part of this research project.

IRB Code Number: 1409P53924
Version date: Oct. 22, 2014

***Online Test Instructions**
You will now start the *MODEST* test. This test includes 12 open-ended questions.

As a reminder, the purpose of this assessment is to evaluate how you think about a problem from a statistical point of view.

Directions:

1. Read the brief article to help you familiarize yourself with the problem scenario that you will be investigating.

2. Answer the questions related to solving the problem.
   - Be sure to provide as much detail in your answers as possible so someone else can follow your thinking.
   - You will be evaluated based on how you describe your thought processes in your answers.
   - Be sure your answers are complete before moving on to the next page. Once you click the next button to go to the next page, you will not be able to go back to the questions on the previous to review or change your answers.
   - You may want to have writing materials available while you are solving the problem.

Appendix F7: Script for Recruiting Field Test CATALST Students

Hello students,

I am here to invite you to participate in a research project that is developing an assessment of statistical thinking, called Modeling To Elicit Statistical Thinking (MODEST).

The development of this instrument is part of my doctoral dissertation in Statistics Education at the University of Minnesota

You will participate by completing *MODEST* online. The assessment consists of 12 open-ended questions and will take 50 to 70 minutes to complete. Your responses will help me understand how students are thinking statistically on the assessment and will help improve the assessment.

The problem may not look like anything you have done before and the problem doesn't have one correct answer. You do not have to review anything prior to taking the assessment.

While you have to complete this as part of your homework grade for this course, you do not have consent to be a part of my research. [Post administration: In addition, if you demonstrate putting in effort into completing the assessment, you will receive up to 2 points extra credit toward your final homework grade.]

You will receive an email from your instructor within the next day or two.

Here are some handouts that will be used on the assessment. If you lose these, there will also be a link for these handouts in the online assessment.

Thank you.

Appendix F8

Online Consent Form for Field Test CATALST Students

This assessment is part of a research project for developing an assessment called *Modeling To Elicit Statistical Thinking (MODEST).* We ask that you read this form carefully and ask any questions you may have before agreeing to be in the study.

This study is being conducted by Laura Le, a doctoral candidate in the Department of Educational Psychology at the University of Minnesota, under the supervision of Dr. Joan Garfield and Dr. Andrew Zieffler.

Background Information:
The aim of this study is to create an assessment of statistical thinking that attempts to measure the change (i.e., pre/post) in students' statistical thinking as a result of an introductory statistics course, at the undergraduate level. Statistical thinking has generally been considered as "thinking like an expert statistician." The assessment utilizes techniques (e.g., type of problem, open-ended questions) that have been suggested as ways of assessing expert-like thinking.

Procedures:
You will complete an online version of the assessment. The assessment consists of 13 open-ended questions and will take 50 to 70 minutes to complete.

Compensation:
There is no compensation for participating in this research study.

Risks and Benefits of Being in the Study:
As with all research, there is a chance that confidentiality could be compromised. However, we are taking precautions to minimize this risk.

The benefit of participating is the opportunity to develop a better understanding of statistics, and of your own statistical thinking capabilities.

Confidentiality:
The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify you as a participant. All research records will be de-identified and stored on a secure server; only the researchers conducting this study will have access to the records.

Voluntary Nature of the Study:
Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with your institution or the course. If you decide to participate, you are free to withdraw at any time without affecting those relationships.

Contacts and Questions:
The researcher conducting this study is Laura Le under the advisement of Dr. Joan Garfield, Ph.D. (Educational Psychology – Statistics Education) and Dr. Andrew Zieffler, Ph.D. (Educational Psychology – Statistics Education). If you are willing to participate or have any

questions you are encouraged to contact me, Laura Le, at free0312@umn.edu.You may also contact my advisor, Dr. Joan Garfield, at jbg@umn.edu.

If you have any questions or concerns regarding the study and would like to talk to someone other than the researchers, you are encouraged to contact the Research Subjects' Advocate line, D528 Mayo, 420 Delaware Street S.E., Minneapolis, Minnesota 55455; telephone 612-625-1650.

Statement of Consent

Please click the circle below if you agree to participate in this research study.

	I have read the above information and I give permission for my responses to assessment items to be included in any analyses, reports, or research presentations made as a part of this research project.

IRB Code Number: 1409P53924
Version date: Oct. 22, 2014

***Online Test Instructions**
You will now start the *MODEST* test. This test includes 13 open-ended questions.

As a reminder, the purpose of this assessment is to evaluate how you think about a problem from a statistical point of view.

Directions:

1. Read the brief article to help you familiarize yourself with the problem scenario that you will be investigating.

2. Answer the questions related to solving the problem.
   - Be sure to provide as much detail in your answers as possible so someone else can follow your thinking.
   - You will be evaluated based on how you describe your thought processes in your answers.
   - Be sure your answers are complete before moving on to the next page. Once you click the next button to go to the next page, you will not be able to go back to the questions on the previous to review or change your answers.
   - You may want to have writing materials available while you are solving the problem

Appendix F9: Email to (Field Test) CATALST Instructors on Administering MODEST

Dear EPsy 3264 Instructor,

Thank you for allowing me to administer the *MODEST* assessment to your students.

Below is the detailed information for administering the online assessment in your statistics course.

First, send the initial email to your students on…

- (Pre administration)…January 22nd or January 26th.
- (Post administration)…May 6th.

Copy and paste the text below the **Initial Student Email** section below into the body of an email. Be sure that the emails are blind carbon copied (BCC).

Second, send a reminder email to your students on…

- (Pre administration)…January 26th or January 29th .
- (Post administration)….May 12th.

Copy and paste the text below the **Reminder Student Email** section below into the body of an email. Again, be sure that the emails are blind carbon copied (BCC).

Third, I will send an email to you on…

- (Pre administration)…January 30th or February 2nd…
- (Post administration)…May 14th…

with the names of students from your class that completed the assessment.

Thank you again!

Sincerely,

Laura Le
Doctoral Candidate
Department of Educational Psychology
Statistics Education

Initial Student Email:

TO: EPSY 3264 students

You are being asked to take the assessment called *Modeling To Elicit Statistical Thinking* (MODEST). This is the Statistical Thinking Test that needs to be completed for…

- (Pre administration)…Homework 1. Do the best you can as you complete this assignment.
- (Post administration)…Homework 15 or Homework 16. In addition to getting credit for your final homework assignment, you can also earn up to 2 points extra credit toward your total homework grade for giving your best effort in your answers to the questions.

To complete the assessment, please click on the following link:

(Pre administration) https://umn.qualtrics.com/SE/?SID=SV_9ts8TkxPUNsG9X7

(Post administration) https://umn.qualtrics.com/SE/?SID=SV_cHHVM51HpzvLEZn

The assessment consists of 13 open-ended questions and will take 50 to 70 minutes to complete. You can use any resources that you want when completing the assessment *except other people*.

The due date for completing…
- (Pre administration)…Homework 1 is [**January 27th or January 30th].**
- (Post administration)…the assessment is **May 13th by 10p.**

If you have any questions about the assessment, please email Laura Le at free0312@umn.edu.

Thank you! Have a wonderful day!

Sincerely,

[Insert your name here]

Reminder Student Email:

TO: EPSY 3264 students

Last week, you received an email asking you to take the assessment called *Modeling To Elicit Statistical Thinking* (MODEST). Here's your reminder email that you need to take this test in order to get credit for…
- (Pre administration)…Homework 1.
- (Post administration)…Homework 15 or Homework 16, if you have not done so already. Also, in addition to getting credit for your final homework assignment, you can also earn up to 2 points extra credit toward your total homework grade for giving your best effort in your answers to the questions.

If you have not had the chance to complete the assessment, please complete…

- (Pre administration)…Homework 1 by class time tomorrow, [**January 27th or January 30th].**
- (Post administration)…this assessment by **May 13th 10p.**

To complete the assessment, please click on the following link:

(Pre administration) https://umn.qualtrics.com/SE/?SID=SV_9ts8TkxPUNsG9X7

(Post administration) https://umn.qualtrics.com/SE/?SID=SV_cHHVM51HpzvLEZn

The assessment consists of 13 open-ended questions and will take 50 to 70 minutes to complete. You can use any resources that you want when completing the assessment *except other people*.

If you have any questions about the assessment, please email Laura Le at free0312@umn.edu.

Thank you! Have a wonderful day!

Sincerely,

[Insert your name here]