HIGH SCHOOL STUDENTS' UNDERSTANDING OF VARIABILITY

By

STEVEN J. FOTI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2017

To Rachel, my parents, my sisters, and my extended family,
for your love and support throughout this entire journey

ACKNOWLEDGMENTS

There are many people to whom I owe my sincerest gratitude for their help in the success of this dissertation. To my faculty members, fellow graduate students, family, and friends: I am forever thankful for all the support and patience you have provided me throughout my studies. This brief "thank you" expresses only a fraction of my appreciation for you all.

First, I would like to thank my entire dissertation committee for everything that you have done both as part of this dissertation process and otherwise. The suggestions and feedback you have provided helped shape this document and my thinking about education research. Your unwavering willingness to help with course material, projects, and research have significantly contributed to my ability to succeed. Furthermore, I would like to thank Dr. Tim Jacobbe. Chairing my committee was surely a taxing endeavor, but I am truly grateful for your support and guidance throughout the process. Your role as my advisor through the graduate program undoubtedly provided me with experiences that have helped me grow as a learner, researcher, and person. Thank you for all that you have done for me.

Next, I would like to thank my fellow graduate students who allowed me to expend their time discussing ideas, research, and countless other topics. Your support, feedback, and companionship all played a significant role in my success. In particular, I would like to thank Catherine Case and Douglas Whitaker for their friendship and collaboration during our nearly parallel graduate journeys. There is no way to sufficiently summarize my gratitude for you both in writing.

There remain countless other individuals who helped shape this dissertation. Please know that I am forever grateful to everyone not listed here for your contributions. Every journey has its heroes, but the trail is blazed by contributions great and small from those who champion the merit of the destination. Thank you all.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

HIGH SCHOOL STUDENTS' UNDERSTANDING OF VARIABILITY

By

Steven J. Foti

August 2017

Chair: Tim Jacobbe
Major: Curriculum and Instruction

Variability is often considered a central concept in the subject of statistics. This dissertation presents a large-scale snapshot of United States high school students' understanding of variability. Over seven hundred secondary students in high-performing districts from 6 different states, with statistics topics included in their high school standards, participated in the study. Data were collected from the participants using existing instrument designed to measure overall conceptual understanding of statistics through multiple-choice (MC) and constructed response (CR) items. These data were collected in the participants' classrooms during one 90-minute or two 45-minute testing sessions.

Responses to CR items were of primary interest to this study. The items were coded based on how they addressed variability, per an existing framework for understanding statistical variation. Using a procedure developed as part of the study, student responses were scored according to whether the response displayed evidence of understanding of variability. Through quantitative methods, this study was able to identify and utilize trends and patterns in response data to (a) evaluate students' understanding of variability and (b) empirically analyze the role of variability in the overall understanding of statistics. Responses showed strong evidence of understanding how to anticipate variability when collecting data through surveys. However,

there was a glaring lack of evidence of understanding the role variability plays in designing studies and analyzing data. Evidence of understanding of variability found in the CR items was a significant predictor of an overall understanding of statistics. The lack of evidence of strong understanding of variability among secondary students in high-performing districts raises concerns about how high school students, in general, understand the concept.

CHAPTER 1
INTRODUCTION

The importance of statistical literacy has been voiced for decades among statisticians and statistics educators (e.g. Kruskal & Wallman, 1982; ASA, 1991; Steen, 2001). Decisions based on data are made daily by political and business leaders, academics, and government officials. Because these data always vary (Cobb & Moore, 1997), an understanding of the concept of statistical variation is crucial to be able to appropriately collect, analyze, and interpret data. Informed citizens should have a sufficient understanding of statistical ideas to be able to critically evaluate these decisions. To ensure our education system is adequately preparing students to be statistically literate, research into student understanding of variability is necessary.

Research on students' understanding of statistical concepts is well represented in the literature, but tends to focus on measures of central tendency such as the mean, median, mode, and expected value (e.g., Batanero, Cobo, & Diaz, 2003; Cruz & Garrett, 2006; Watson & Moritz, 1999; Shaughnessy & Zawojewski, 1999). This trend is likely due to the heavier emphasis on measures of center in the K-12 curriculum over measures of variability (Shaughnessy, 1997). While research on students' understanding of variability exists, it has the tendency to be limited in scope (e.g., Ben-Zvi, 2004; Reading & Shaughnessy, 2000; Reading, 2004; Shaughnessy & Ciancetta, 2002; Torok & Watson, 2000). For instance, the results of most studies focused on variability tend to analyze very small samples of students. While these studies reveal a great deal about how individual students think about variability, they do not allow for the recognition of patterns and trends across students. Research that analyzes a larger number of student responses to a wide variety of tasks is needed to paint a broader picture of students' current understanding of variability.

The use of statistics assessment items that focus on variability is one way to collect data that displays students' understanding of variability. Prior research adapted items from the National Assessment of Educational Progress (NAEP) assessment to develop interview tasks that provided insight into students' understanding of variability (e.g. Shaughnessy et al. 1999; Reading & Shaughnessy, 2000). However, assessments such as NAEP have been criticized for assessing statistical understanding using items that focus on procedural understanding of statistics rather than conceptual understanding. The NSF-funded Levels of Conceptual Understanding of Statistics (LOCUS) project (DRL-1118168) was launched, in part, to change the way statistics is assessed (Jacobbe et al., 2014). LOCUS resulted in a set of new assessments that measure conceptual understanding of statistics (Jacobbe, 2015) in a manner that is consistent with the K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) framework (Franklin et al., 2007) while also addressing state and national standards (e.g., Common Core State Standards). The GAISE framework identifies variability in data as a defining characteristic of the discipline of statistics and emphasizes its importance throughout the statistical problem-solving process. Thus, the LOCUS assessments can be used to explore students' understanding of variability in greater depth.

This dissertation seeks to provide a large-scale snapshot of high school students' understanding of variability by using an existing framework that describes robust understanding of variability (Peters, 2011). In particular, this study examines students that have been taught some amount of statistics and come from schools in high-performing districts according to standardized testing. Proportions of students that understand elements of variability will be examined from two different perspectives, design and data-centric, to determine which elements are well understood by high school students and which may require more attention in the

curriculum. The design perspective deals with the acknowledgement, recognition, and/or anticipation of variation in the design of a study. The data-centric perspective considers variation that is represented, measured, and described during exploratory data analysis (Peters, 2011). Additionally, a regression model will be used to determine the relationship between an understanding of variability and an overall conceptual understanding of statistics. While not free from limitations, this study will provide a broader image of how high school students with some statistics instruction are understanding the concept of variability.

Variability and Statistical Literacy

Variability is omnipresent in data that are used to make decisions in daily life and is frequently cited as one of the main reasons that the discipline of statistics exists (Cobb & Moore, 1997; Moore & Cobb, 2000; delMas, 2004). Understanding the role variability plays in data analysis is necessary for members of society to both make and critically evaluate data-based claims and findings (Gal, 2004). While statistics is commonly seen as a subfield of mathematics, many statisticians would argue that it is a mathematical science (Cobb & Moore, 1997) used to solve problems in the presence of variability. Statistics is more similar to fields such as economics and physics in which mathematical tools play a large role but do not represent the essence of the field (Cobb & Moore, 1997; delMas, 2004). Each of these fields has its own core concepts that guide the exploration of information. In the case of statistics, variability is one of those core concepts.

Mathematics content standards and curriculum recommendations reflect the importance of variability in grades 6-12 statistics education. The National Council of Teachers of Mathematics (2006) explicitly emphasizes measures of variability in curriculum recommendations for statistics and data analysis as early as Grade 6. The *Principles and Standards* (NCTM, 2000) strand for data analysis at the high school level is centered around

students “drawing conclusions in light of variability” (p.325). The influence of the Common Core State Standards for Mathematics (CCSSM) (NGACBBP & CCSSO, 2010) has further increased the role of statistics in mathematics curricula across the United States. Beginning in Grade 6, students are expected to develop an understanding of statistical variability by anticipating its existence when collecting data and describing measures of spread in a dataset. By high school, students should be prepared to fully explore the effects of variability when making predictions and decisions with data. The increased recognition of variability as a central concept to statistics in curriculum recommendations and content standards creates a need for further research on student understanding.

The concept of variability is important for high school students to understand because it is a fundamental aspect of how practicing statisticians approach empirical inquiry (Wild & Pfannkuch, 1999). Empirical inquiry requires a thorough understanding of the investigative cycle, which is synonymous to the statistical problem-solving process, and ideas considered to be fundamental to statistics (Wild & Pfannkuch, 1999). These fundamental concepts include recognizing the need for data, transnumeration, consideration of variability, reasoning with statistical models, and integrating statistical and contextual information. Statistical reasoning and thinking are used to describe the skills and knowledge required to thoroughly conduct investigations using data (Ben-zvi & Garfield, 2004). They can also be thought of as the skills and thought processes utilized by practicing statisticians (Pfannkuch & Wild, 2004). Variability can be considered a core concept in statistical thinking because it is the motivating idea behind the data collection, analysis, and interpretation phases of empirical inquiry (Moore, 1992; 1990).

Additionally, variability is important for high school students because it is a central component of how citizens interpret and make conclusions about data (Wallman, 1993; Steen,

2001; Gal, 2002; Franklin et al., 2007). The ability to effectively consume statistical information and data is known as statistical literacy, which has a growing focus in some K-12 statistics education recommendations (Franklin et al., 2007). Statistical literacy is defined in multiple ways throughout the statistics education literature (e.g., Wallman, 1993; Watson, 1997; Garfield, 1999; Gal, 2000; Franklin et al., 2007). In the most general sense, statistical literacy refers to the ability to "intelligently cope with the requirements of citizenship, employment, and family, and to be prepared for a healthy, happy, and productive life" in a data-driven world (Franklin et al., 2007, p. 1). This dissertation more specifically uses the term to describe people's ability to interpret and critically evaluate information and data-based arguments that appear in diverse media channels (Gal, 2000). The implied meaning of a statistically literate person is one that has developed a basic understanding of all four components of the statistical problem-solving process—formulate questions, collect data, analyze data, and interpret data—and their underlying concepts.

To achieve statistical literacy, students should understand the roles that variability plays in statistical investigations (Moore, 1990; Cobb & Moore, 1997; Shaughnessy, 1997; Moore, 1998; Garfield & Gal, 1999; Gal, 2004; Franklin et al., 2007). Per the K-12 GAISE framework (Franklin et al., 2007), formulating a statistical question requires the ability to anticipate variability in the data collected in order to answer the question. If a posed question has a deterministic answer, then it is not considered a statistical question. Collecting data requires the ability to acknowledge variation for the purpose of controlling potential sources of variability. These techniques help reduce the amount of variability in the data and ensure the conclusions of the study are meaningful. Statistical techniques for analyzing data are used to give an accounting of the variability in the data. Margins of error, confidence intervals, and standard deviations all

use the distribution of repeated sampling to account for variability in the sample data. Data reduction techniques, such as graphing and calculating summary statistics, are used to find the key features and trends that are often hidden by variation in the data (Wild & Pfannkuch, 1999; Konold & Pollatsek, 2002). Finally, making interpretations about data in the presence of variability requires the allowance of variability. Generalizations must be made that extend beyond the data and allow for variability in the sample responses. The natural existence of variability in data and its importance in the statistical problem-solving process requires educators and researchers to carefully consider the emphasis on variability in teaching and learning. Determining how students are currently understanding the concept is a necessary step in this consideration.

Theoretical Underpinnings

Consideration of variation is believed by many to be a requirement of the ability to think statistically (e.g., Franklin et al., 2007; Shaughnessy, 1997; Wild & Pfannkuch, 1999) and is often a goal of introductory statistics courses (e.g., Ben-Zvi & Garfield, 2004; Chance, 2002). Thus, research on students' reasoning about and understanding of variability is necessary to better determine how to teach statistics in ways that emphasize the concept of variability. This dissertation draws on the theoretical perspective employed in existing research regarding students' understanding of variability in order to add to the knowledge base.

Originally conceptualized by Piaget (1954, 1962), cognitive developmental models attempt to describe the changes and dynamics in how people understand mathematics and other domains. Piaget hypothesizes that learners develop knowledge through a series of stages that are tied to context-neutral and biologically driven universal structures. However, Piaget also acknowledges a constructivist aspect of learning that recognizes the influence of the environment and of educational intervention. These two ideas are somewhat contradictory to each other and

have led to the rise of neo-Piagetian cognitive developmental theorists that replace the universal stage theory with domain-specific theories (e.g., Bidell & Fischer, 1992; Biggs & Collis, 1982, 1991; Case, 1985; Case & Okamoto, 1996; Fischer, 1980). Biggs and Collis (1982), for example, built off of Piaget's work to develop the Structure of Observed Learning Outcomes (SOLO) model, which focuses on students' responses instead of their cognitive level of development.

Many of these models are utilized in research that examines students' mathematical reasoning and thinking in areas such as geometry, number operations, and probability. This dissertation specifically draws from the Biggs and Collis (1982, 1991) model because of its use in cognitive developmental models in students' statistical reasoning (e.g., Jones et al., 2000; Mooney, 2002; Watson et al., 1995). The original SOLO taxonomy (Biggs & Collis, 1982) consists of five modes of functioning—sensorimotor, ikonic, concrete-symbolic, formal, and post-formal—and five cognitive levels—prestructural, unistructural, multistructural, relational, and extended abstract—that exist within each mode and represent increased complexity in students' understanding. Later additions to the SOLO model (Biggs, 1989; Biggs & Collis, 1991; Collis & Biggs, 1991; Pegg & Davey, 1998) acknowledge that students' development in earlier modes supports development in later modes.

The SOLO-based Framework for Robust Understanding of Statistical Variation (Peters, 2011) is utilized in this study for its thorough descriptions of specific elements of variation from different perspectives. The design perspective deals with the acknowledgement, recognition, and/or anticipation of variation in a study design. Anticipating and acknowledging variability in the design perspective are synonymous to the roles variability plays in formulating questions and collecting data, as described in the K-12 GAISE framework (Franklin et al., 2007). The data-centric perspective considers variation that is represented, measured, and described during

exploratory data analysis (Peters, 2011). This perspective is closely related to the analyzing data component of the statistical problem-solving process from the K-12 GAISE framework, which is used to account for variability in a collected dataset.

The Framework for Robust Understanding of Statistical Variation (Peters, 2011) also identifies four elements of variation that transcend the perspectives: variational disposition, variability in data for contextual variables, variability and relationships among data and variables, and effects of sample size on variability. Reasoning with a variational disposition involves anticipating and acknowledging variability when considering study design, data collection techniques, analysis of data, and interpretations derived from data. Reasoning about variability in data for contextual variables incorporates the context of the data in both the anticipation of variability in study design and the consideration of potential sources of variability in data. The element of variability and relationships among data and variables deals with controlling variability through study design and exploring controlled and random variability in data. Finally, the effects of sample size on variability was identified as an overarching element of understanding. These four elements are considered to define integrated reasoning of variability across the different perspectives (Peters, 2011).

The theoretical underpinnings of this dissertation draw from cognitive developmental learning theory and, more specifically, the SOLO model. These theories, in conjunction with frameworks for statistical reasoning and the understanding of statistical concepts that resulted from them, provide the basis for this study on students' understanding of variability. The structure of student responses can be examined to determine their current level of understanding of statistical variation and the extent of their ability to make connections between elements of variation.

Research Questions

This dissertation aims to provide a large-scale snapshot of high school students' understanding of variability who have been taught some amount of statistics and attend schools in high-performing districts. The Framework for Robust Understanding of Variation (Peters, 2011) will be used to describe the design and data-centric perspectives of variation and the four overarching elements of variability. Students' understanding will be examined through their responses to items from the LOCUS assessments. Because the LOCUS assessments were developed to measure overall conceptual understanding of statistics (Jacobbe, 2015), this study will also investigate the relationships between understanding of variability and overall understanding of statistics. The following research questions will guide the study:

1. What proportion of high school students understand variability from the design and the data-centric perspectives, and each of the four overarching elements of variability? Do high school students score higher on items from a particular perspective?
2. What is the relationship between overall conceptual understanding of statistics and understanding of variability from the design and data-centric perspectives amongst high school students?

Working Definitions

Many terms used throughout this dissertation are not universally defined in statistics education literature. Some terms have been defined above, and the rest of the terms are defined here for clarification for the reader. Unless otherwise indicated, the working definitions below are the intended meanings throughout the dissertation:

1. Variation: used to describe the act of varying or changing condition.
2. Variability: used interchangeably with variation, as is consistent with literature (see Peters, 2011; Garfield & Ben-Zvi, 2008).
3. Statistical Literacy: people's ability to interpret and critically evaluate information and data-based arguments that appear in diverse media channels (Gal, 2000).

4. Conceptual Understanding: knowledge that is rich in depth and makes connections between statistical ideas and concepts.

5. Procedural Understanding: a familiarity with the procedures or formulas used to answer questions but a lack of deeper knowledge.

Structure of the Dissertation

The written portion of this dissertation will include five chapters. This chapter (Chapter 1) serves as an introduction and contains an overview of the relevant literature and a theoretical framework, which together provide the context for the study. Chapter 2 contains a review of relevant literature that is necessary to ground and justify this study. Chapter 3 explains the methodology used in this study, including a description of the participants and instruments used. Additionally, the procedures for the development of a scoring procedure and analysis of the quantitative data collected will be discussed along with the limitations of this study. The results of the study will be presented in Chapter 4. Chapter 5 will serve as the concluding chapter of the dissertation, and will discuss the implications and limitations of this research as well as related future research ideas.

CHAPTER 2
REVIEW OF THE LITERATURE

As a result of the importance of the concept of variability to statistical literacy, multiple research studies have been conducted to gain insight into how students understand it (e.g., Shaughnessy et al., 1999; Reading & Shaughnessy, 2000; Torok & Watson, 2000; Reading, 2004; Reid & Reading, 2008). These research studies tend to use a specific task, or set of tasks, to construct empirically supported frameworks to determine where students are in their developmental understanding of variability. The following chapter of this dissertation will review statistics education literature that focuses on students' understanding of variability. Since studies on this topic scarcely focus primarily on secondary students, this review will include research conducted with K-12 and tertiary students as well. These studies will be examined to summarize findings, shed light on the development of frameworks for the understanding of variability, and provide a literature base for the current study.

Student Understanding of Variability

Research on the student understanding of variation was rather scarce prior to 1999 (Ben-Zvi, 2004). Since then, there has been an increase in research on students ranging from grades 4-12 as well as tertiary students (e.g. Torok & Watson, 2000; Reading, 2004; Reid & Reading, 2008). To help organize the review of this literature and better orient the review with this dissertation, Peters' (2011) framework for robust understanding of variation will be employed for its description of variation from different perspectives. The modeling perspective integrates reasoning used to fit models to patterns of variability in data and to determine how well these models fit. The data-centric perspective integrates reasoning about exploring, measuring, and representing variation in the analysis of data. These two perspectives were adapted and extended by Peters (2011) for use with understanding of variability from research on understanding of

distributions (Prodromou & Pratt, 2006). The adoption of perspectives from understanding of distributions is supported by literature that recognizes variability as an integral piece of understanding distributions (Reading & Reid, 2006). Finally, the design perspective captures reasoning about variation that anticipates and acknowledges variation in the design of quantitative studies. The purpose of including the design perspective is to capture requisite reasoning about variation (Wild & Pfannkuch, 1999) that is not accounted for by the other two perspectives. Since the design and data-centric perspectives are the primary focus of this dissertation, and few studies specifically target student understanding of variation from the modeling perspective, the review prioritizes studies from the former two perspectives.

Design-Perspective

The design perspective is the perspective that includes much of the research, which is reasonable due to its focus on the basics of study design. Furthermore, a significant line of research involves the use of the lollies problem—a sampling task with lollies that was used as the main focus in multiple studies *(*Shaughnessy et al., 1999; Reading & Shaughnessy, 2000; Torok & Watson, 2000). The lollies problem is an adaptation of a 1996 National Assessment of Educational Progress (NAEP) item that researchers used to learn more about student understanding of variability in the setting of a sampling task. Results from the 1996 NAEP administration showed that no students commented on the issue of spread (Shaughnessy & Zawojewski, 1999). However, researchers felt that redesigning the task would encourage students to reveal their understanding of variability.

Other research that falls under the design perspective includes student understanding of variability in a probability scenario (e.g. Shaughnessy & Ciancetta, 2002). The types of questions regarding variability are like those asked in the sampling situations. For example, in a sampling situation, one might ask the number of times an item would be expected to be chosen in each of

10 trials, and in a probability situation, one might ask the number of times a particular outcome would be expected to occur in each of 10 trials. Both questions reveal similar information about student understanding of variability, however, changing the task from sampling to one that involves probability may yield different results.

Shaughnessy et al. (1999) used the adaptation of the 1996 NAEP item in a study of grades 4-12 students in Australian and American schools. The written responses revealed that students did not have a strong understanding of the center or spread of outcomes from the sampling task, like the responses from the original NAEP administration. One trend was that grade 12 students seemed to provide responses that reported low estimates of variation, possibly due to instruction in probability and lack of instruction in sampling distributions. Students also failed to use specific words like "vary," "deviate," or "variation" when attempting to indicate variation. To further explore what students were thinking, the researchers decided to interview students in grades 4-12 using a protocol involving the lollies problem (Reading & Shaughnessy, 2000).

Reading and Shaughnessy (2000) interviewed twelve students from Australian schools and reported the results from four of those interviews, one in each of grades 4, 6, 9, and 12. The students were given two different sampling scenarios. The first involved a bowl of 100 lollies where 20 of them were yellow, 50 were red, and 30 were blue, and the second also involved a bowl of 100 lollies where 20 were yellow, 70 were red, and 10 were blue. In each of the scenarios, students were asked to consider how many red lollies they would expect to get in a handful of 10 lollies. Additionally, they were asked to consider how many red lollies they would expect each of six people to get if they took a handful of 10 lollies. The students had an actual

bowl of colored lollies that contained the same proportions of each color as the scenario presented and could revise their responses after having an opportunity to conduct the activity.

The condensed student response form is shown in Figure 2-1. Note that in addition to a description of the sampling scenario, the researchers included prompts that specifically address variability. For example, the question labeled 1A in Figure 2-1 concludes with, "would this happen every time? Why?" to determine students' anticipation of variability in various samples drawn from the bowl. Students were prompted to respond in three different ways: list, choice, and range. The first way required the student to list the number of reds they expected each of the six people to draw in their handful of 10 lollies. The second way required students to select which of the choices presented would best represent the number of reds each of the six people would draw. Finally, the students were asked to give a range of likely number of reds drawn from lowest to highest.

Students' responses to the form in Figure 2-1 were coded to understand how they perceived center and spread in the sampling situation. If a response was consistent with the expected value of reds that would be drawn in a sample of 10 lollies, it was coded as mean-centered. If a response was below the expected value it was coded as low and if it was above the expected value it was coded as high. For spread, responses were coded as narrow, reasonable, or wide according to the dispersion. The purpose of the interviews and coding scheme was to compare students' responses across the three versions of the task and to examine their perceptions of variability.

The study was exploratory in nature, and did not specify any theoretical framework about student understanding of variability. However, the interviews and coding of written responses

were used by researchers to make insightful observations that drew from their expertise in statistics and statistics education.

Student Response Form

1A) Suppose we have a bowl with 100 lollies in it. 20 are yellow, 50 are red, and 30 are blue. Suppose you pick out 10 lollies. How many reds do you expect to get? __ Would this happen every time? Why?

1B) Altogether six of you do this experiment. What do you think is likely to occur for the numbers of red lollies that are written down? Please write them here. _____, _____, _____, _____, _____, _____ Why are these likely numbers for the reds?

1C) Look at these possibilities that some students have written down for the numbers they thought likely. Which one of these lists do you think best describes what might happen? Circle it.

a) 5,9,7,6,8,7

b) 3,7,5,8,5,4

c) 5,5,5,5,5,5

d) 2,3,4,3,4,4

e) 7,7,7,7,7,7

f) 3,0,9,2,8,5

g) 10,10,10,10,10,10

Why do you think the list you chose best describes what might happen?

1D) Suppose that 6 students did the experiment—pulled out ten lollies from this bowl, wrote down the number of reds, put them back, mixed them up. What do you think the numbers will most likely go from? From ____ (low) to ____ (high) number of reds. Why do you think this?

** (After doing the experiment) Would you make any changes to your answers in 1B–1D? If so, write the changes here.

Figure 2-1. Student response form (condensed) (Shaughnessy et al., 1999).

The researchers discovered that while students showed growth in their ability to describe the sampling situation on the centering scale, there was no consistent growth in their ability to

describe the situation on the spread scale. The 12th grade student gave narrow responses for the spread, and students were generally not able to use specific language to explain their reasoning. Both tendencies were confirmed by previous research with the scenario (Shaughnessy et al., 1999). Listing responses tended to provide more information about the students' implicit conception of variation than the choice and range forms of response.

In a later study, Torok and Watson (2000) aimed to expand on the research done by Shaughnessy et al. (1999) by conducting interviews that would more deeply probe student understanding of variability and investigate their knowledge of terms associated with variability. They conducted sixteen interviews with 8 boys and 8 girls from a school in Tasmania, where there were two students from each of grades 4, 6, 8, and 10. The researchers chose to use four different situations that involved variation: two involving the lollies problem, one involving real daily local temperature data, and one involving the heights of a large group of children. The goal was to use a combination of scenarios with isolated random variables, like the lollies problem, and scenarios with real world variation, that consisted of multiple sources of variation.

The researchers analyzed the interviews to develop a four-level hierarchical framework that described levels of developing concepts of variation. A student at level A displayed a "weak appreciation of variation," at level B an "isolated appreciation of … variation," at level C an "inconsistent appreciation of variation," and at level D a "good appreciation of variation" (Torok & Watson, 2000). Like the study by Shaughnessy et al. (1999), the small sample size limited the conclusions that could be drawn from the results. However, the researchers could use their framework to describe characteristics of student understanding of variability based on those analyzed in the sample. For example, students at levels A and B demonstrated an affinity to individual values or sub-ranges when predicting the outcomes of the sampling situations where

students at levels C and D could describe their expected outcome as a middle value with other likely outcomes surrounding it. Somewhat surprisingly, little association seemed to exist between understanding of variability and real-world knowledge. Students' responses to the questions that involved heights and temperature, scenarios that students would be more familiar with, did not show different levels of understanding than their responses to the questions involving the lollies problem, a scenario that they were less likely to have experience with.

In another study, Reading and Shaughnessy (2004) qualitatively analyzed all the interview data collected by Shaughnessy et al. (1999), Shaughnessy and Zawojewski (1999), and Reading and Shaughnessy (2000) for the lollies problem, not just the four case studies that were analyzed by Reading and Shaughnessy (2000). The researchers utilized the hierarchical framework developed by Torok and Watson (2000) as a starting point to analyze the student responses. After analyzing the responses, the authors felt that two additional hierarchies were needed to describe certain characteristics of the responses—one based on how students described the variation and one based on how the students explained the source of the variation. One important finding resulting from the use of these additional hierarchies was that the form of response (list, choice, range) influenced whether the student attempted to describe the variance or look for a cause of it.

Conclusions drawn from these studies were not meant to be descriptive of a population, but were intended to be starting points for further research regarding understanding of variation in a random sampling setting. Reading and Shaughnessy (2004) pointed out that the time-consuming process of analyzing qualitative interview data restricted the sample size of these studies, but provided richer information than just analyzing written responses. The task itself was also limiting because reasoning about variation occurred in many other types of scenarios other

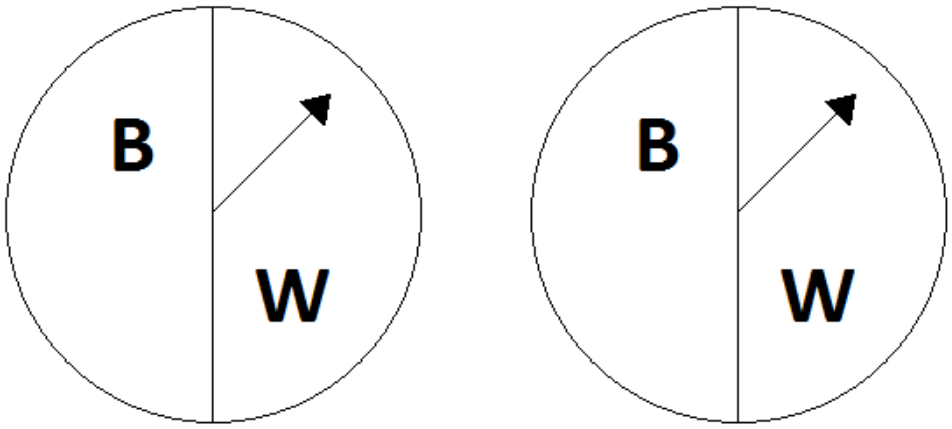
than a sampling situation with isolated random variation and a known population. Additionally, having students only consider 6 random samples from the bowl of lollies was noted to be a potential limitation to students' ability to consider the variation that occurs in repeated sampling (Reading & Shaughnessy, 2000). For example, having them respond with the number of reds that would be expected in each of 100 samples, possibly through simulations, may have led students to more thoughtfully consider variation. When having students choose from a list of possible outcomes, researchers later considered allowing students to describe all options that they felt made sense, which would yield more information about their thinking than forcing them to choose one option. In some cases, students that were responding to the surveys lacked sophisticated language when discussing variation, which may have limited their ability to communicate their thoughts clearly (Shaughnessy et al., 1999; Reading & Shaughnessy, 2000).

The studies using the lollies problem revealed how little students understood about variability and shed some light on potential problems with mathematics and statistics curricula. In general, students tended to show growth in their understanding of measures of central tendency as they got older (from grade 4 to 12), however, no such growth appeared for understanding spread (Shaughnessy et al., 1999; Reading & Shaughnessy, 2000; Torok & Watson, 2000). The success with measures of central tendency may have been a result of the K-12 curriculum's focus on finding the mean, median, mode, and expected values, and limited focus on measures or interpretations of spread or variability (Shaughnessy, 1997). Torok and Watson (2000) linked increased understanding of central tendency *and* variation with increases in grade level. Differences in conclusions may have been because Torok and Watson did not have Grade 12 students in their study, and they suggest further research in Grade 12 students' understanding is needed. Reading and Shaughnessy found that Grade 12 students tended to select

responses that overemphasize the expected value of the sampling situation (e.g. "5, 5, 5, 5, 5, 5"). The researchers theorized that this was a systematic bias due to Grade 12 students having formal instruction in probability and not in sampling distributions (Shaughnessy et al., 1999; Reading & Shaughnessy, 2000).

Shaughnessy and Ciancetta (2002) conducted a study that described student understanding of variability within the design perspective, but used a probability task instead of a sampling one. Another 1996 NAEP item, shown in Figure 2-2, was used to generate discussion about variation that involved two identical spinners that had one black side and one white side. The results from the NAEP item showed that most students were unable to provide the correct answer with correct reasoning. For example, only 28% of 12th grade students in the general population provided correct answers, and only 8% of the 12th grade students also provided correct reasoning (Shaughnessy & Zawojewski, 1999).

Shaughnessy and Ciancetta wanted to change the item to be more statistical in nature and wanted to see how students taking various mathematics classes responded across grade levels. The original, unchanged task, as shown in Figure 2-2, was administered to 652 students in grades 6-12 taking 28 different math classes in 5 different schools. Results showed that most students in grades 6-8 answered incorrectly by stating that the chances of winning were 50-50. Some growth was shown for students in grades 9-12 algebra courses, but still only between 30% and 43% of students answered correctly for these classes. Students in upper level courses such as pre-calculus, AP Calculus, and AP Statistics did quite well on the item, each achieving around 90% correct responses with correct reasoning. The researchers hypothesized that playing the game and seeing the variability in outcomes would help students that struggled with this item to recognize the sample space.



The two fair spinners above are part of a carnival game. A player wins a prize only when *both* arrows land on black after each spinner has been spun once. Jeff thinks he has a 50-50 chance of winning. Do you agree?

A Yes B No Justify your answer.

Figure 2-2. Spinner task from 1996 National Assessment of Educational Progress

To address their claim, they interviewed 28 students in grades 8-12 with similar demographic profiles as the students that answered the survey item. During the interview, the students could respond to the original item, explain their answer, and then were asked how many times they would expect someone to win the game if they played 10 times. After running actual trials with the spinner, the students had the opportunity to reconsider their answer to the original item. Interviews after students had the opportunity to play the game revealed that the number of students who believed the game was not 50-50 went up and the number who exhibited correct reasoning went up. Additionally, seeing the variability in repeated trials led more students to list the sample space on their own. Researchers concluded that having students conduct simulations that reveal variation in the outcomes of a task helps them make connections to the sample space. However, because some students still insisted that the game was 50-50 after collecting data, often with the stance that "anything can happen when you play, the game should be 50-50," the

researchers agreed with previous research that states it can be very difficult to change existing student beliefs about probability and statistics (Shaughnessy, 1992).

The studies that involved student understanding of variability from the design perspective revealed that students tended to have trouble with anticipating variability. Having to imagine what potential outcomes might occur in a hypothetical sampling situation or a probability situation is another example where students are required to think abstractly. However, some of the research provides evidence that allowing students to conduct the trials that are explained in the scenarios helps students understand the variability in the outcomes (e.g. Shaughnessy & Ciancetta, 2002). Other interesting challenges arise because of these studies, including the effects of instruction in probability on student understanding of the concept of variability. As Cobb and Moore (1997) discussed, formal probability is a difficult concept to master, and is not necessary for beginning courses in statistics. These studies provide evidence that instruction in formal probability may negatively interfere with students' ability to reason in scenarios that involve a probabilistic element but do not require formal knowledge of probability.

Additionally, results from these studies are not necessarily generalizable to any larger population of students. They all examined small samples of students in a qualitative manner. The results might only be specific to the students interviewed, and they might also include bias from the researchers. While further studies are needed to continue to understand the errors and misconceptions in reasoning in situations that involve variability, there is growing evidence that many students have difficulty with the concept from the design perspective.

Data-Centric Perspective

The data-centric perspective considers variability that appears in a set of data. Studies that fall under this perspective include one where students respond to an open-ended task about weather data to make inferences (Reading, 2004), one that focuses on student understanding of

variability that can be seen in data displays (Meletiou & Lee, 2002a), and one that has students examine the effect of variability when comparing two groups (Ben-Zvi, 2004). These studies involve different types of tasks than those seen in the design perspective, however, many of the analysis methods and results are closely related.

Reading (2004) aimed to extend the research on student understanding of variability by applying and, if necessary, expanding the hierarchy used by Reading and Shaughnessy (2004), which was developed in a sampling context, to a context that involved a natural event in which variation occurred. One class from each of grades 7, 9, and 11 in a secondary school in Australia were selected to participate in the study. The researchers had the students work on a weather activity that used real rainfall and temperature data to help students write a report on which month would be best to hold an outdoor celebration. Data from 36 months was used to ensure that each student had a different dataset to analyze. Each student was given the chance to draw conclusions from the data individually before getting together with a small group to decide on which month they all thought was best for the celebration. The task was intentionally left open-ended to allow students to talk about variability whenever they felt it was important.

Researchers analyzed individual student responses using a three-step approach consistent with previous research (e.g. Mooney, 2002; Langrall & Mooney, 2002; Watson et al., 2003). They began by coding the responses according the hierarchy used by Reading and Shaughnessy (2004). Then, descriptors were revised based on any new descriptions of variation, and the hierarchy was expanded for responses that the existing framework did not explain well. All responses were then coded using the new hierarchy. Researchers discovered that while most responses fit within the levels of the existing hierarchy, some responses described variation using words while others described variation using only numbers. Thus, they determined that these two

types of responses were different enough to warrant an extension of the original framework to include two developmental cycles, one for qualitative responses and one for quantitative responses. The responses showed that beginning conceptions of variability may start with word descriptions, but as students develop their understanding of variability, they begin to use numbers to describe the variability.

Students' qualitative responses were sorted into two categories: limiting and sequential. The limiting responses set a general limit on the values in the data. For example, a limiting response about the weather data might suggest that a month is not a good choice because it is too cold or too hot. A sequential response deals with the data item by item or by blocking parts of the data in a qualitative way. For example, a sequential response about rainfall might state that there were a few dry days and then a couple of wet days. The quantitative responses, which were assumed to be more statistically sophisticated, were sorted according to whether the response referred to exterior values, interior values, or both in the data. For example, stating the minimum and the maximum showed that the response was focused on exterior values, where stating the interquartile range would imply that the response was focused on interior values. In comparing their hierarchy to the one developed by Reading and Shaughnessy (2004), the researchers noticed that none of the students' responses discussed variability in a manner that focused on deviations from an anchor. The study revealed that students were not responding to the task completely as intended, and that very few higher-level conceptions of variability were represented in the responses.

The intent of the real-world context was to create a meaningful and engaging task for the students to work with that included real data that varied naturally. However, there was strong evidence that the nature of the task limited students' ability to interpret variation in the data. For

example, the use of real data seemed to interfere with students using techniques they recently learned in class like making visual displays of the data. Additionally, student motivation on the task began to drop off when some realized that the celebration described in the task was only hypothetical. However, researchers maintained the high quality of the information gleaned from the study, especially because they were not attempting to quantify the best students but to discover what types of descriptions were used to address variability in a dataset.

Another study that falls under the data-centric perspective dealt with comparing two groups using real world data. Ben-Zvi (2004) aimed to examine how students began to reason about variability during a group-based task given in a supportive environment. Two Grade 7 students who were considered to have above average ability and verbal skills were videotaped and interviewed during and after they worked together to complete the surnames task. The task consisted of 35 Hebrew surnames collected from their class, 35 English surnames from an American class, and the length of each surname. The main task was presented by the classroom teacher and consisted of comparing the surname length of the two groups.

The focus of the analysis was on how the students began to "notice and acknowledge variability in the data and make use of special local information in different ways as stepping-stones towards the development of global points of view of describing and explaining the variability between groups" (p. 47). The study identified seven stages of development that the students progressed through on their path to a conclusion. The initial focus of students was on individual pieces or local features of the data, such as noticing how many names included "Mc," and they eventually worked up to more global features of the data, such as how long names usually were in each group. The researchers noted that the students' development of reasoning

about variability happened simultaneously with their development of seeing global features of the distributions as wholes by noticing shape, center, and spread.

Some factors that appeared to have helped the students develop their reasoning about variability were noted by the researchers. For example, the students used multiple informal tools and methods in their attempts to understand the variability in the data, the students had previous experience with exploring data, and the students' experience with the context of the data allowed them to make connections between their statistical observations and their knowledge of the context. Further, student interactions with the teacher acted as a catalyst to their thinking. They would occasionally prompt the teacher for feedback which, although limited in nature, would help them move forward in their thinking. Ben-Zvi considered the learning to have taken place in a zone of proximal development (Vygotsky, 1978), where the teacher acted as the expert that absorbed the students' understandings into their own framework and then provided feedback for the students to review their thinking and create new understanding.

Because of previous research that implied students had trouble thinking stochastically, Meletiou and Lee (2002b) aimed to develop a statistics course that put variability at the center of everything. The idea was that by developing students' understanding of variability, they would more deeply understand other concepts in statistics. The authors developed an introductory statistics course that was taught by the second author at a university in the United States to a class of 33 students in business-related fields. The instructor attempted to increase student awareness of variation through a series of investigations that highlighted the purpose of statistics as a set of tools and methods that are necessary to handle variation.

As part of the course, graphical displays were introduced during investigations as tools used to visualize trends, patterns, and deviations from those patterns. One of the primary types of graphical displays used in the course was the histogram. Assessment data collected from students at the beginning of the course showed that very few could decide, with correct reasoning, which of two distributions, shown with histograms, had more variability. An open-ended questionnaire at the end of the course combined with interviews of eight students provided evidence that the course had influenced students' statistical thinking to be less deterministic. However, students still struggled with certain concepts, such as sampling distributions, not only because of the abstractness of the idea, but also because of their still limited understanding of histograms.

The studies that took place in the data-centric perspective continued to reveal information about how students describe and understand variation. In some cases, researchers' understanding of student descriptions of variation were expanded, for example when Reading (2004) discovered that description hierarchies and causation hierarchies were missing from the existing framework used in a sampling situation. Additionally, it was revealed that students may begin to reason with variation in a much less statistically sophisticated way using non-specific words and phrases to describe what occurs in the data. As students progressed in their understanding, they began to use more numerical information to support their ideas with statistical measures of variation. This phenomenon was not only seen in the weather task (Reading, 2004), but also in the surnames task (Ben-Zvi, 2004) in which the two students began to use numbers from the data to help describe the comparison between the two groups.

These studies not only revealed information about student understanding of variability from a different perspective but also revealed new challenges that exist in different contexts. Contrary to the findings of Torok and Watson (2000), the studies in the data-centric perspective

provided evidence that the context did play a role in the students' experiences. In the surnames task, students appeared to utilize their knowledge of the context to assist in their statistical reasoning (Ben-Zvi, 2004). However, negative effects appeared when Reading (2004) began to see lack of motivation because students realized that the context, which involved planning a celebration, was not actually going to occur. Also, because of the students' familiarity with the context, they seemed to ignore skills and tools learned in the classroom, for example displaying the data graphically, when attempting to glean information from the data.

Development of Frameworks for Student Understanding of Variability

Introduction and First Frameworks

Early research in student understanding of variability used student responses to NAEP items adapted to a survey or brief interview protocol (Shaughnessy et al., 1999, Reading & Shaughnessy, 2000). To analyze the students' responses, researchers used their expertise as statisticians and statistics educators to make conclusions about the students in their studies. No explicit frameworks were employed to describe student understanding in either of these studies nor was there explicit mention of any frameworks used in analyzing the data.

Torok & Watson (2000) were the first to explicitly describe their process for analyzing qualitative data in a study regarding student understanding of variation. In addition to the lollies problem, researchers also used two other scenarios that involved a set of temperature data and the heights of a large group of children. Sixteen 45-minute interviews were conducted with boys and girls in grades 4, 6, 8, and 10 from two different schools in Australia and responses were transcribed for analysis. To analyze the data, the researchers used a clustering technique like the one described by Miles and Huberman (1994) and previously used in other research in statistics education not directly related to variation (Watson & Moritz, 2000; Mokros & Russel, 1995).

Level A: Weak appreciation of variation
• Acknowledge variation
• Provide responses that suggest a very weak understanding of proportional ideas
• Focus on individual outcomes without consideration of the set
• May refer to the average as the most common individual value
• Provide answers with inconsistent degrees of variation and clustering
• Are easily swayed by experimental results
• Do not produce meaningful summary graphs (for 40 draws).
• Never refer to variation explicitly, show poor knowledge of variation terminology
• Have poor general knowledge of real world situations

Level B: Isolated appreciation of aspects of variation and clustering
• Readily acknowledge variation
• Provide responses that suggest a very weak understanding of proportional ideas
• May provide answers in terms of sub-ranges or specific values
• May refer to the average as a value within a range of common values
• May provide answers with consistently too much or too little variation and clustering
• Are moderately swayed by experimental results
• Generally attempt summary graphs but do not produce meaningful ones (for 40 draws)
• Never refer to variation explicitly, have reasonable knowledge of variation terminology
• Have variable knowledge of real world situations

Level C: Inconsistent appreciation of variation and clustering
• Readily acknowledge variation
• Exhibit strong proportional thinking and may provide responses that imply representativeness, such as the "perfect" sample of 5 red, 2 yellow, and 3 green
• Provide answers in terms of specific outcomes in the context of a set of outcomes
• May provide answers with consistently too much or too little variation and clustering
• Are only slightly influenced by experimental results
• Produce equivalent of time series graphs to summarize data
• Explicitly refer to variation, may have strong knowledge of variation terminology
• Have basic general knowledge of real world situations

Level D: Good, consistent appreciation of variation and clustering
• Readily acknowledge variation
• Provide responses that suggest a conflict between proportional ideas; or exhibit strong proportional thinking and provide responses that imply representativeness.
• Provide answers as specific outcomes in the context of a set of outcomes
• Consistently provide answers with an appropriate level of clustering
• Are only slightly influenced by experimental results
• Produce frequency or time series graph to summarize data
• Explicitly refer to variation, usually have strong knowledge of variation terminology
• Have good general knowledge of real world situations

Figure 2-3. Four levels of developing concepts of variation (Torok & Watson, 2000).

First, a checklist of important statistical concepts, ideas associated with the interview protocol, and initial observations from the student transcripts was created using previous literature as a guide (Moore, 1990; Shaughnessy, 1997; Shaughnessy et al., 1999). The checklist was finalized after revisions and refinement resulting from the re-reading of interview transcripts and consisted of four general groupings of understanding of variation. The four groups were: sensitivity to variation in relation to likelihood of individual outcomes, sensitivity to variation in relation to distribution of consecutive outcomes, language used during interviews, and aspects related to real world contexts. Each main group consisted of between 2 and 4 subgroups that further described the main group. Each student response was coded either 0, 1, or 2 for each subgroup, where 0 represented no or poor demonstration of understanding and 2 represented complete or successful demonstration.

These coded responses were used to conduct a visual search for clustering in the data matrix and the resulting clusters were used to describe four levels of understanding of variation: no appreciation, weak appreciation, inconsistent appreciation, and good/consistent appreciation of variation and clustering. The four levels were further described using examples from the students' responses. The framework, shown in Figure 2-3, represents tentative themes in students' responses to these questions, and the researchers call for further studies to help build the bigger picture of how a larger sample of students respond to these types of questions about variability.

As Reading and Shaughnessy (2004) analyzed qualitative data from the lollies problem using the framework for student understanding of variability developed by Torok and Watson (2000), they discovered that the existing framework did not adequately account for all the students' responses. As discussed above, they extended the framework to include a description

hierarchy to explain how students describe the variation that occurs, and a causation hierarchy to describe how students attempt to explain the source of variation. While not explicitly using any theoretical framework for cognitive growth, the description and causation hierarchies each include four levels that represent increasing sophistication in student responses. Torok and Watson's framework focused mainly on evidence that students were appreciating or recognizing the variability in the task, where the description and causation hierarchies are more specific to how the students describe and explain variation. For example, the description hierarchy, shown in Figure 2-4, moves from responses that discuss either middle or extreme values, labeled D1, to responses that discuss values as deviating from a central anchor, labeled D4. The description hierarchy is much more directly applicable to student responses because it contains identifiable features from the response, whereas the Torok and Watson hierarchy may involve more interpretation by the reader.

Levels	Focus of Responses
D1 - Concern with Either Middle Values or Extreme Values	Describe variation in terms of what is happening with either extreme values or middle values. *Extreme Values* are used to indicate data items that are at the uppermost or lowest end of the data, while *Middle Values* indicate those data items that are between the extremes.
D2 - Concern with Both Middle Values and Extreme Values	Describe variation using both the extreme values and what is happening with the values between the extremes.
D3 - Discuss Deviations from an Anchor	Describe variation in terms of deviations from some value but either the anchor for such deviations is not central, or not specifically identified as central.
D4 - Discuss Deviations from a Central Anchor	Describe variation by considering both a center and what is happening about that center.

Figure 2-4. Description of Variation Hierarchy (Reading & Shaughnessy, 2004) as displayed in Reading (2004).

Levels	Descriptors
No consideration of variation	Do not display any meaningful consideration of variation in context Do not acknowledge variation in relation to other concepts (e.g., distribution)
Weak consideration of variation	Identify features of only one source of variation (within-group or between-group) Acknowledge variation in relation to other concepts Incorrectly describe variation Do not base description of variation on the data Anticipate unreasonable amount of variation Poorly express description of variation Refer to irrelevant factors to explain variation incorrectly refer to relevant factors to explain variation Do not use variation to support inference
Developing consideration of variation	Clearly describe both within-group and between-group variation Recognize the effect of a change in variation in relation to other concepts Correctly describe variation Base description of variation on the data Anticipate reasonable amount of variation Clearly express description of variation Correctly refer to relevant factors to explain variation Use variation to support inference Do not link the within-group and between-group variation
Strong consideration of variation	Link within-group and between-group variation to support inference

Figure 2-5. Consideration of Variation Hierarchy (Reid & Reading, 2008)

Because there was little research on student understanding of variability at the tertiary level, Reid and Reading (2004) further refined an existing hierarchy used at the pre-tertiary level. During a one-semester introductory statistics course for students in science-related fields in an Australian university, 46 students completed minute papers that focused on the curricular topics of exploratory data analysis, probability, sampling distributions, and inferential statistics. The authors used one of the five types of thinking described by Wild and Pfannkuch (1999), consideration of variation, and the hierarchy for student understanding of variability developed

by Torok and Watson (2000) in order to build a hierarchy for student consideration of variability. Their framework consisted of four levels—no, weak, developing, and strong consideration of variation—with descriptors that explained features of the students' responses to each of the four minute problems. The authors used a four-item pre- and post-test questionnaire (Reid & Reading, 2005), a class test, and assignments with the hierarchy developed for the minute papers in order to create a combined hierarchy of student consideration of variation, shown in Figure 2-5 (Reid & Reading, 2008).

The resulting hierarchy is intended to be used for students responding to tasks related to a college-level introductory statistics course. To assess student consideration of variation in more advanced statistical tasks, more research is needed to refine the descriptors at each level. The researchers caution that because consideration of variation is a complex problem, students should not be assessed based solely on a response to a single task. Additionally, the researchers found very few responses that were coded as displaying a strong consideration of variation. Thus, further extension of the framework is necessary to develop descriptors for the highest level of consideration.

The SOLO Taxonomy in Studies about Variation

The SOLO model (Biggs & Collis, 1991) has been employed in probability (e.g. Jones et al., 1997) and in statistics (e.g. Jones et al., 2000; Mooney, 2002) to create and describe developmental hierarchies. Most of the studies in this review that utilize the SOLO model operate in the ikonic and concrete symbolic modes, which represent making use of imaging and operating with second order symbols, respectively.

The framework for statistical thinking for young children developed by Jones et al. (2000), and later refined for middle school students by Mooney (2002), was one of the first in the statistics education literature to employ the SOLO taxonomy. The framework consisted of four

SOLO-based levels in each of four processes. Only the one process, organizing and reducing data, is relevant to student understanding of variability and is shown below in Figure 2-6. The four levels explain responses that move from the idiosyncratic level, where students are not able to describe the spread of data in terms that are relevant to discussion about spread, to the analytical level, where students can describe the spread of the data using correct and valid measures of spread. This framework was often used as a guideline for the development of future frameworks based on the SOLO taxonomy.

	Organizing and Reducing Data
Levels	**Focus of Responses**
1 - Idiosyncratic	Is not able to describe the spread of the data in terms representative of the spread.
2 - Transitional	Describes the spread of the data using invented measures that are partially valid.
3 - Quantitative	Describes the spread of the data using a measure from a flawed procedure or a valid and correct invented measure.
4 - Analytical	Describes the spread of the data using a valid and correct measure.

Figure 2-6. Statistical thinking framework (Mooney, 2002) as displayed in Reading (2004).

Watson et al. (2003) developed a questionnaire with items that were designed to allow students to demonstrate their understanding of variability in sampling contexts, like the lollies problem, and data and chance contexts, which are the statistics strands most students see in school settings. The authors used the SOLO taxonomy to code the student responses to the questionnaire in a hierarchical fashion. Some items were coded with a three-point hierarchical scale, others with a four-point scale, and a few with five- and six-point scales. The coded data

was used in conjunction with a Rasch partial credit model to develop an empirically supported framework for understanding of variability.

The quantitative analysis of the coded data revealed four levels of understanding of variation, shown in Figure 2-7: prerequisites for variation, partial recognition of variation, applications of variation, and critical aspects of variation. The thresholds for the levels were based on judgements about the "content, sophistication, and structure of the responses" (p. 13). The results of the coding and the levels of understanding implied that there was a continuum of understanding and students were often floating around the border of two different levels. The Rasch analysis, along with evidence of strong content validity, implied that all the items, despite their various contexts, measured a unidimensional construct that the authors identified to be variation. This result was in line with many researchers' beliefs that variation is at the center of statistical investigation (e.g. Shaughnessy, 1997; Moore, 1990; Cobb & Moore, 1997; Wild & Pfannkuch, 1999).

Levels	**Focus of Responses**
1 - Prerequisites for variation	Working out the environment, table/simple graph reading, and intuitive reasoning for chance.
2 - Partial recognition of variation	Putting ideas in context, tendency to focus on single aspects and neglect others
3 - Application of variation	Consolidating and using ideas in context, inconsistent in picking most salient features.
4 - Critical aspects of variation	Employing complex justification or critical reasoning.

Figure 2-7. Developing Concepts of Variation (Watson et al., 2003) as displayed in Reading (2004).

Although the Watson et al. (2003) framework was developed to measure understanding of variation, the four levels do not explain how students explicitly describe variation. Reading and Shaughnessy (2004) expanded on the Torok and Watson (2000) framework to establish a hierarchical framework that would do just that. Because of the depth of information that the SOLO taxonomy allows in analyzing students' responses, Reading (2004) used the final version of the Reading and Shaughnessy (2004) framework to not only see how it would hold up in a data-based inference setting, but also see if SOLO could be used as a conceptual framework to explain the hierarchical model.

While coding the student responses, Reading (2004) found that the original hierarchy did not adequately explain all the responses. Because many students were describing the variation using only words in a less sophisticated way, Reading added a second level to the description hierarchy that accounted for qualitative responses. The descriptors remained nearly identical to the ones used for D1 and D2 in the original framework (Figure 2-4), however the qualitative versions were less sophisticated. There were no students in the study that displayed responses at the D3 or D4 levels, which correspond to discussing deviations from an anchor or central anchor, respectively. Thus, Reading suggests that further research explores these areas to determine if that part of the framework can also be extended. However, Reading identified a third level above D1 and D2 responses that not only considered both middle, or interior, and extreme values, but also showed the ability to link the two sets of values.

The SOLO taxonomy was used to explain the cognitive growth throughout the six levels of the model, shown in Figure 2-8. In the qualitative cycle of the framework, the element of interest in this study was determined to be "a feature of the variation of the data described qualitatively" (Reading, 2004, p. 98). A unistructural response was one that contained one such

element, a multistructural response contained more than one such element, and a relational response linked these elements. Many of the qualitative responses were intuitive notions that were not necessarily straightforward to interpret. Researchers suggested looking at other research to help understand these notions. For example, Makar and Confrey (2003) found that pre-service teachers often used unclear or informal terms to compare dot plots but could more thoroughly explain their thinking. These explanations could be useful in helping to define the intuitive notions that younger students have and add to the descriptors at each level of the framework.

First Cycle Qualitative Responses	**Element - qualitative feature of variation of data**
U1 - unistructural	Magnitude related or arrangement related
M1 - multistructural	Limiting related and/or sequential related
R1 - relational	Link the general limit with discussion of blocks sequentially
Second Cycle Quantitative Responses	**Element - quantitative feature of variation of data**
U2 - unistructural	Based on extreme values or interior values
M2 - multistructural	Based on extreme values and/or interior values
R2 - relational	Linking of extreme values and interior values may suggest immature notions of deviations

Figure 2-8. Refined Description of Variation Hierarchy (Reading, 2004)

In the second cycle of the model, the quantitative cycle, the element of interest was "a feature of variation in the data described quantitatively" (Reading, 2004, p. 99). Again, a unistructural response contained one such element, multistructural responses contained more than one, and relational responses linked multiple elements. Research, such as the study conducted by Lann and Falk (2003) to evaluate strategies used by statistically naive tertiary

students to compare data for greater variability, was again used to help identify certain characteristics of students' responses. While the range, for example, clearly deals with extreme values and the interquartile range deals with the interior values, students also use measures that are not as clearly extreme- or interior-value focused. Lann and Falk attempted to analyze student justifications for their responses, but this proved to be a difficult task. Results, however, would likely prove to be more useful for defining the qualitative cycle of the model.

Watson et al. (2007) later conducted a study that combined many of the previously discussed tasks to identify a hierarchy for, and consider developmental pathways between, students' understanding of expectation and variation. Their study consisted of 73 students in grades 1, 3, 5, 7, and 9 in Australian schools. Five protocols were used to garner student responses to the lollies task (Shaughnessy et al., 1999), the weather task (Torok & Watson, 2000), a task comparing two distributions of student test scores (Watson & Moritz, 1999), the spinner task (Shaughnessy & Ciancetta, 2002), and a population/sample means task (Tversky & Kahneman, 1971). The researchers used the SOLO taxonomy as a structure for their analysis. Coding schemes for each of the tasks were adapted versions of those used by the researchers who originally analyzed their respective data from the tasks. Using a Rasch analysis, Watson et al. (2007) identified a 6-level hierarchical framework for student understanding of expectation and variation, shown in Figure 2-9. Using the resulting model and a review of previous literature, the researchers hypothesized pathways between understanding of expectation and variation and how they develop throughout the 6 levels.

This framework appears to be consistent with the frameworks previously developed under each of the specific tasks used in this study. For example, the Reading (2004) framework explains how students describe variability in terms of the students' use of words and/or numbers

to make connections between extreme and interior values. The Watson et al. (2007) framework considers establishing links between expectation and variation which, in different words, is related to the idea of explaining some notion of center and the amount of change or deviation from that center. The authors stress that while it may be difficult to interview students in order to utilize the framework in a classroom setting, the individual activities could easily be adapted for small group work or discussion to collect evidence for student understanding of variability. Additionally, the authors suggest that the framework is useful for determining what types of teacher interventions are appropriate. Teachers can pinpoint a student's level of understanding by using level descriptors and then target activities to help them consolidate "necessary underpinning knowledge" which leads them into the next level (p. 115).

Levels	**Descriptors**
1 - Idiosyncratic	Little or no appreciation of either expectation or variation.
2 - Informal	Primitive or single aspects of expectation and/or variation and no interaction of the two.
3 - Inconsistent	Acknowledgement of expectation and variation, often with support, but few links between them.
4 - Consistent	Appreciation of both expectation and variation with the beginning of acknowledged interaction between them.
5 - Distributional	Established links between proportional expectation and variation in a single setting.
6 - Comparative Distributional	Established links between expectation and variation in comparative settings with proportional reasoning.

Figure 2-9. Framework for Conceptual Understanding of Expectation and Variation (Watson et al., 2007).

Although there have been multiple frameworks developed in the last two decades to explain understanding of variability, Peters (2011) argued that there was still no "holistic image of robust understanding of variation" (p. 52). The existing frameworks, while insightful and informative, failed to reach beyond the concrete-symbolic mode of cognitive functioning because they all describe understanding of variability in ways that are directly tied to their context. Peters wanted to develop a framework that would thoroughly describe a robust understanding of variation in the formal mode of functioning, or reasoning that both incorporates and transcends the context of interest.

Using the SOLO taxonomy as their supporting theory, Peters (2011) hypothesized a framework that included two UMR (unistructural, multistructural, relational) cycles of levels of response. In the first cycle, three perspectives of variation are presented: the design perspective, the data-centric perspective, and the modelling perspective. In the second cycle, robust understanding of variation is described as reasoning across all the perspectives. The elements of each perspective and descriptors for those elements were developed by analyzing a pilot study of six secondary mathematics/statistics teacher-leaders' (Peters, 2009) and a main study of 16 teacher-leaders' responses to three main tasks that corresponded with the three perspectives. The resulting framework, shown in Figure 2-10, included four main elements for each of the three perspectives: variational disposition, variability in data for contextual variables, variability and relationships among data and variables, and effects of sample size on variability. Further, each of these elements contained multiple, more specific indicators from the analysis of the teachers' responses.

Elements and Reasoning Indicative of Robust Understanding of Variation

Perspective / Element	Design Perspective	Data-centric Perspective	Modeling Perspective
Variational disposition	DP1: Acknowledging the existence of variability and the need for study design	DCP1: Anticipating reasonable variability in data	MP1: Anticipating and allowing for reasonable variability in data when using models
Variability in data for contextual variables	DP2: Using context to consider sources and types of variability to inform study design or to critique study design	DCP2: Describing and measuring variability in data for contextual variables as part of exploratory data analysis	MP2: Identifying the pattern of variability in data or the expected pattern of variability for contextual variables
Variability and relationships among data and variables	DP3: Controlling variability when designing studies or critiquing the extent to which variability was controlled in studies	DCP3: Exploring controlled and random variability to infer relationships among data and variables	MP3: Modeling controlled or random variability in data, transformed data, or sample statistics
Effects of sample size on variability	DP4: Anticipating the effects of sample size when designing a study or critiquing a study design	DCP4: Examining the effects of sample size through the creation, use, or interpretation of data-based graphical or numerical representations	MP4: Anticipating the effects of sample size on the variability of a sampling distribution

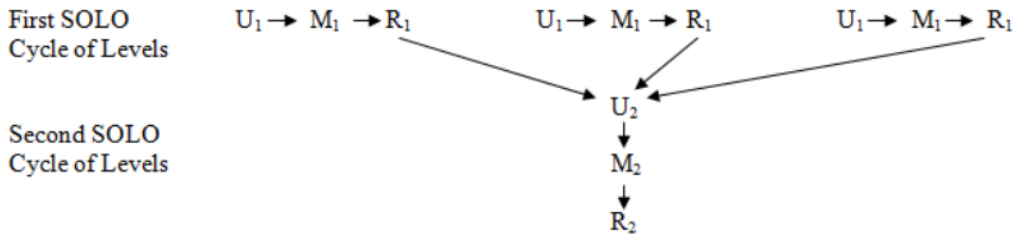


Figure 2-10. Framework for Robust Understanding of Variation (Peters, 2011)

This framework was the first to provide a holistic image of a robust understanding of variation in the formal mode. While other frameworks provide information about understanding of variation, they are often tied completely to the context and do little to answer calls to examine reasoning beyond concrete-symbolic reasoning (e.g. Reading, 2004). This framework not only answers these calls, but provides deep descriptions of within-perspective reasoning and across-perspective reasoning. It provides researchers, teachers, and curriculum developers with a goal of instruction and pathways to improve understanding of variation. The framework not only provides the opportunity to analyze student responses to individual tasks on a level that transcends those tasks, but also allows researchers to analyze classroom discourse to determine

whether the discourse supports the development of a robust understanding of variation. Finally, the framework is a useful tool to inform teacher educators and future teachers of statistics of the kind of robust understanding of variation an expert teacher is expected to have.

Discussion

The SOLO model has proved to be a useful tool in the development of frameworks that explain student understanding of variability. It allows researchers to deeply analyze student responses to develop an empirically-based framework (Reading, 2004). For example, Reading (2004) used an existing framework, that was not developed under the SOLO taxonomy, to code new student response data. The results of the coding showed that the original framework was not detailed enough to fully explain the range of responses, and required an additional level to account for qualitative responses. The SOLO taxonomy supported this finding because it allows for multiple cycles of cognitive growth. Within each cycle, the SOLO taxonomy also explained the development from D1 to D2 and beyond as unistructural, multistructural, and relational. Another strength of frameworks developed using the SOLO taxonomy is that they provide a basis for creating rubrics (Reading, 2004; Reid & Reading, 2008; Peters, 2011). Assessing student understanding of variation by using level descriptors from the frameworks can give an idea of where a student is in their cognitive development of the concept. Additionally, these frameworks can be used to inform teaching by examining students' views of variation, determining their level of understanding, and using the level descriptors to design interventions to help improve their understanding.

The development of these frameworks relies solely on drawing meaning from student responses. In some cases, these are written responses to tasks, while other cases include interview protocols. In either case, the understanding of variation displayed in the response is entirely determined by the interpretation of the researchers. The studies discussed above used

multiple coders, and any coding discrepancies were resolved through discussion until there was complete agreement. Still, it can be difficult to interpret student responses when they are unclear or use informal terminology (Reading, 2004). However, more research on teachers' understanding of statistical concepts could help alleviate this problem because they may be able to more aptly describe their informal understanding (e.g. Makar & Confrey, 2003). Another limitation of these frameworks is that, because they attempt to describe cognitive development, it is necessary to have students respond to multiple tasks to get a clear picture of their understanding. Students often have issues with contexts which may bolster or undermine their ability to express their understanding (e.g. Torok & Watson, 2000; Reading, 2004).

The previously discussed studies paint a slightly confusing picture about how students are understanding variability. Despite the growth of statistics in the K-12 curriculum, students at the secondary and tertiary level still show a general lack of understanding of concepts that are widely considered to be central to the discipline of statistics. Not only do students have difficulty expressing their thoughts using language native to statistics, but even their informal conceptions lack evidence of a solid understanding of variability. In multiple instances, researchers developed frameworks that included levels that were above the level of reasoning of most of the participants in the study (e.g. Reading, 2004). This lets us know that many students are facing challenges with reasoning about variability and are unable to reason at a level that one would expect to see after instruction in statistics. The research shows many possible explanations for this lack of advanced reasoning such as confusion due to similarity with the context or even due to weak instruction in probability. Clearly, the concept of variability needs to take a more central role in statistics courses, perhaps following the example of the course designed by Meletiou and Lee (2002b).

This general lack of a solid understanding of variability appears in other related studies as well. Research on student understanding of more advanced statistical concepts such as distribution traces errors and misconceptions back to a weak understanding of variability (Reading & Reid, 2006; Pfannkuch & Reading, 2006). Another topic of study is students' understanding of measures of variability, such as the standard deviation (e.g., delMas & Liu, 2005). Initially, students will be introduced to variability via the standard deviation. Although the measure of standard deviation is somewhat complex and confusing to novice students, the idea of variability can be taught in simpler ways. In one study, Mathews and Clark (2003) discovered that students who earned a grade of A in an introductory statistics course could not explain the standard deviation beyond its computation. Research about student understanding of samples and sampling distributions also revealed that student understanding of variation played a major role in their reasoning (Watson, 2004; Chance et al., 2004).

However, before drastic changes to the landscape of statistics education are made, consideration of the types of studies that have been reviewed is necessary. Many of the studies discussed above are qualitative in nature and draw from small sample sizes. While these qualitative studies are very detailed and informative for individual cases, they do not necessarily reveal widespread patterns across many students. Additionally, the coding procedures used in these studies rely on interpretations by the lead author and their team of researchers. Experience and the agreement among multiple independent coders provide credibility, but do not rule out misinterpretation of what a student means in a response. In many cases, the researchers called for further investigation of student understanding in their respective tasks to help develop the framework. Unfortunately, framework development seems to be short lived, as researchers turn to new and different tasks that explore variability. Instead, more research needs to start with

general frameworks, like the one developed by Peters (2011), and collect more data to test and refine the framework. This dissertation seeks to add to the knowledge base of student understanding of variability by analyzing responses to tasks from constructed response items through the lens of the framework created by Peters.

CHAPTER 3
METHODOLOGY

Introduction

A major goal of this study was to describe U.S. high school students' understanding of variability to inform future research, teaching practice, and curriculum in statistics. Previous studies focusing on students' understanding of variability have used small samples sizes to develop in-depth reports on how students think about variability (Ben-Zvi, 2004; Reading & Shaughnessy, 2000; Reading, 2004; Shaughnessy & Ciancetta, 2002; Torok & Watson, 2000). To provide a large-scale snapshot of student understanding, this study analyzed data from a large sample of high school students. Student response data on an assessment of conceptual understanding of statistics was examined using quantitative methods to focus on overarching patterns in their understanding of variability from different perspectives. This study can contribute to future discussions about the role of variability in statistics curricula and which perspectives may require more attention. In this chapter, I review the research questions that guided this investigation and the design of the study used to answer them. Specifically, I describe the process used to select participants, collect data, and analyze the data to explore students' understanding of variability.

Theoretical Perspective

This dissertation analyzed scored student responses to items on the Levels of Conceptual Understanding of Statistics (LOCUS) assessments through the perspective of neo-Piagetian cognitive developmental theory, and will use the Structure of Observed Learning Outcomes (SOLO) model (Biggs & Collis, 1982, 1991) as a tool to examine students' understanding of variability. Neo-piagetian cognitive developmental theory and the SOLO model acknowledge that students' understanding increases in complexity as they mature, and allow for different

levels of complexity based on the domain. This study specifically considered U.S. high school students' understanding of variability across a few different contexts from LOCUS and shed light on which elements and perspectives of variability need more attention.

The theoretical framework underpinning this study supports the notion that students' understandings of the concept of variability may differ from their level of understanding of mathematics or their level of understanding of statistics as a whole. Prior research on student's understanding of variability (e.g., Shaughnessy et al., 1999; Reading & Shaughnessy, 2000; Torok & Watson, 2000) shows that many students exhibit low levels of understanding when faced with tasks involving variability. These studies utilized long-form tasks, follow-up interviews, and context specific frameworks. The use of LOCUS to understand students' understanding of variability would provide a more streamlined method of assessment for potential use in the classroom setting. To examine student understanding of variability that transcends the specific context of the problem, the Robust Understanding of Statistical Variation framework (Peters, 2011) was employed because it was developed in the formal mode of cognitive functioning.

Study Design

Research Questions

To analyze high school students' understanding of variability using the LOCUS assessments, this study used the Robust Understanding of Statistical Variation framework (Peters, 2011) as a guide. The two-dimensional framework considers four elements of variability from three different perspectives and includes descriptions of what robust understanding looks like for each point of intersection between element and perspective (Table 3-1).

Table 3-1. Descriptors for elements and perspectives of variability in the Robust Understanding of Variability framework (Peters, 2011).

	Design Perspective	Data-Centric Perspective	Modeling Perspective
Variational disposition	DP1: Acknowledging the existence of variability and the need for study design in a. controlling the effects of variation from extraneous variable(s); b. including considerations of variation for variable(s) of interest during data analysis; or c. using sample statistics to infer population parameters for the variable(s) of interest	DCP1: Anticipating reasonable variability in data by a. considering the context of data; b. recognizing that data descriptions should include descriptions or measures of variability (and center); or c. recognizing unreasonable variability in data (e.g., that which could result from a data entry error)	MP1: Anticipating and allowing for reasonable variability in data when using models for a. making predictions from data; or b. making inferences from data
Variability in data for contextual variables	DP2: Using context to consider sources and types of variability to (1) inform study design or to (2) critique study design by a. considering the nature of variability in data (e.g., measurement variability, natural variability, induced variability, and sampling variability); or b. anticipating and identifying potential sources of variability	DCP2: Describing and measuring variability in data for contextual variables as part of exploratory data analysis by a. (1) creating, (2) using, (3) interpreting, or (4) fluently moving among various data representations to highlight patterns in variability; b. focusing on aggregate or holistic features of data to describe variability in data; or c. (1) calculating, (2) using, or (3) interpreting appropriate summary measures for variability in data (e.g., measures of variation such as range, interquartile range, standard deviation for univariate data sets; correlation and coefficient of determination for bivariate data sets)	MP2: Identifying the pattern of variability for contextual variables by a. modeling data to explain variability in data; b. considering contextual variables in the formulation of appropriate data models; c. considering contextual variables in modeling data to describe holistic features of data; or d. considering or creating distribution-free models or simulations to explore contextual variables

Table 3-1. Continued.

	Design Perspective	Data-Centric Perspective	Modeling Perspective
Variability and relationships among data and variables	DP3: Controlling variability when (1) designing studies or (2) critiquing the extent to which variability was controlled in studies by a. using random assignment or random selection of experimental/ observational units to (in theory) equally distribute the effects of uncontrollable or unidentified sources of variability; or b. using study design to control the effects of extraneous variables (e.g., by incorporating blocking in experimental design or stratifying in sampling designs) to isolate the characteristics of the variable(s) of interest or to isolate systematic variation from random variation	DCP3: Exploring controlled and random variability to infer relationships among data and variables by a. (1) using and (2) interpreting patterns of variability in various representations of data; b. focusing on aggregate or holistic features of variability in data to make comparisons; c. (1) using or (2) interpreting appropriate summary measures of the variability in data to make comparisons (e.g., transformed versus untransformed data); or d. examining the variability within and among groups	MP3: Modeling controlled or random variability in data, transformed data, or sample statistics for a. making inferences from data (e.g, isolating the signal from the noise for univariate or bivariate sets of data or formally testing for homogeneity in variances); or b. assessing the goodness of a model's fit by examining the deviations from the model
Effects of sample size on variability	DP4: Anticipating the effects of sample size on the variability of a. a sample or b. statistics used to characterize a sample (e.g., mean, proportion, median) when (1) designing a study or (2) critiquing a study design	DCP4: Examining the effects of sample size on the variability of a. a sample or b. statistics used to characterize a sample (e.g., mean, proportion, median) through the creation, use, or interpretation of data-based graphical or numerical representations	MP4: Anticipating the effects of sample size on the variability of a sampling distribution to a. model the sampling distribution; or b. consider significance, practical or statistical significance, of inferences

The following research questions were investigated to develop a snapshot of U.S. high school students' understanding of variability:

1. What proportion of U.S. high school students understand variability from the design and the data-centric perspectives, and each of the two overarching elements of variability? Do U.S. high school students score higher on items from a particular perspective?
2. What is the relationship between overall conceptual understanding of statistics and understanding of variability from the design and data-centric perspectives amongst U.S. high school students?

Both research questions were answered using quantitative methods to utilize a large amount of assessment response data and draw overarching conclusions regarding U.S. high school students' understanding of variability. The first question was answered by scoring student responses to LOCUS CR items per the understanding of variability they displayed. The procedure that was used for scoring the responses is described in the data analysis section of this chapter. Tabulated scores were used to determine the proportion of students that showed evidence of understanding of variability for cross-sections of element and perspective as well as specific elements and perspectives. The second question used a multiple linear regression model to determine the importance of the concept of variability to overall conceptual understanding of statistics, as measured by the multiple-choice section of the LOCUS assessment. The regression model controlled for various demographic features of the participants in an attempt to isolate the relationship between variability and overall understanding of statistics.

Instrument

The LOCUS assessments were the result of an NSF-funded project (DRL-1118168) to develop an instrument that measured conceptual understanding of statistics. They were developed using a modified version of Mislevy and Riconscente's (2006) evidence-centered design (ECD) process, which involved the construction of assessments through evidentiary arguments. The five layers of ECD, domain analysis, domain modeling, conceptual assessment

framework, assessment implementation, and assessment delivery, were utilized in the creation of an evidence model that served as a blueprint throughout the development of the assessment (Jacobbe et al., 2015). The description of conceptual understanding of statistics in the K-12 GAISE framework (Franklin et al., 2007) served as the foundation for the evidence model. The evidence model contains descriptions for each element of the problem-solving process at every developmental level, evidence statements of understanding the element, possible work products, and observable features of responses demonstrating understanding and served as the foundation for item writing.

The GAISE framework (Franklin et al., 2007) describes understanding of statistics using a two-dimensional model focusing on the components of the statistical problem-solving process and developmental levels. The three developmental levels, A, B, and C, are based on development in statistical literacy. Understanding of each component of the problem-solving process—formulating questions, collecting data, analyzing data, and interpreting data—is described across the three developmental levels. At level A, a person should begin to develop data sense and an understanding of basic statistical tools. The nature of variability at this stage is limited to measurement, natural, and induced variability that occurs within groups of interest. At level B, a person should continue to build on concepts from level A and begin to see statistical reasoning as a way to solve problems using data. Also, sampling variability is introduced and the focus of variability is extended to both within and between group variation. At level C, a person should be able to understand the statistical solving process at a deep enough level to explain statistical reasoning to others. They should be comfortable formulating statistical questions, appropriately collecting data to answer the questions, and analyzing and interpreting the data to

form conclusions. Random variation and the role it plays in the inference process is also explored at this level.

The item writing process for LOCUS resulted in a set of multiple choice items that connected directly to elements in the evidence model, and constructed response items coded by developmental level and process component. Two versions of the LOCUS assessment were constructed using the written items, a Beginning/Intermediate version and Intermediate/Advanced version. The Beginning/Intermediate version of the assessment corresponds with levels A and B of the GAISE framework, and the Intermediate/Advanced version corresponds with levels B and C. Within each version of the assessment, there are also two equated forms that are available to be used in a pretest/posttest manner or as standalone assessments.

Pilot studies for the LOCUS assessment items were conducted with students in grades 6-12 that had some form of statistics instruction. Of the 3324 students that participated in the pilot study, 2075 completed the Beginning/Intermediate and 1249 completed the Intermediate/Advanced versions of the assessment. Nearly all (95%) of the high school students that participated in the pilot study completed the Intermediate/Advanced version. Results from the pilot study revealed that many students struggled with the material (Jacobbe et al., 2015), but high school students tended to construct higher quality responses on CR items (Foti & Jacobbe, 2015; Whitaker & Jacobbe, 2014, Case & Jacobbe, 2014). This study chose to use the Intermediate/Advanced version of LOCUS to focus on high school students, where higher quality responses were expected.

The two equated forms of the Intermediate/Advanced version share two out of five CR items in common. The other three items are different in context, but cover the same process

components at identical developmental levels. Form 2 was chosen for use in this study because more of the CR items addressed variability from the design and data-centric perspective than those on form 1. This version of the LOCUS assessment consisted of 23 MC and 5 CR items. The reliability, estimated with stratified alpha, for form 2 of the Intermediate/Advanced version of LOCUS was 0.87 and has been validated as a measure of conceptual understanding of statistics for high school students (see Jacobbe et al., 2014; Jacobbe et al., 2015).

Participant Selection

The participants (N = 742) in this study were secondary students from schools in Florida, New Jersey, Arizona, Colorado, Ohio, and Georgia. These states were chosen based on three criteria: a contact for the LOCUS project was in the area for ease of implementation, the school was in a high-performing district according to standardized assessment scores, and the state standards included statistics prior to the CCSSM. The students had all taken, or were in the process of taking, a statistics course or a mathematics course that included statistics, and were in Grades 9 (n = 6), 10 (n = 105), 11 (n = 211), or 12 (n = 421). Participants were nearly evenly split between male (46%) and female (51%), with roughly 3% omitting to respond to the gender survey item. A majority of the students were White (62%), non-Hispanic (82%), and reported that English was the primary language spoken at home (79%). A full table of the participants' demographics is shown in Table 3-2.

By visually comparing the participants in this study with the approximate distribution of demographics among all secondary students in the United States (NCES, 2014), some clear differences can be seen. Participants selected for this study from high-performing districts in states that had statistics in their state standards tended to include more White, Asian, and non-Hispanic students than in the general population. Whether English was the primary language spoken at home was not available for the entire United States secondary school population,

however, among the population 5 years of age and older, about 20% primarily speak a language other than English at home (ACS, 2015).

Table 3-2. Demographics of students in sample and approximate percentages of secondary students in the U.S.

	N	%	U.S. %
Gender			
Female	382	51.41	49
Male	342	46.03	51
Omitted	19	2.56	
Grade			
9	6	0.81	
10	105	14.13	
11	211	28.40	
12	421	56.66	
Ethnicity			
Not Hispanic	610	82.10	74.60
Mexican	53	7.13	
Puerto Rican	16	2.15	
Cuban	12	1.62	
Other Hispanic	30	4.04	
Omitted	22	2.96	
Race			
Native American	10	1.35	1.0
Asian	45	6.06	5.3
Black	108	14.54	15.50
White	460	61.91	49.50
Other	70	9.42	
Omitted	50	6.73	
English Spoken			
No	129	17.63	
Yes	588	79.14	
Omitted	26	3.50	

Students in this study were in schools from six different districts around the United States. Part of the reason these districts were chosen to participate in the LOCUS project was because they were considered high-performing districts (Jacobbe et al., 2015). Each school was

instructed to administer LOCUS to students in high level courses, including honors level mathematics, AP Statistics, and AP Calculus, in an attempt to gather higher quality responses.

Within a participating school, all students in eligible classrooms took the assessment. However, only data from students whose parents signed an informed consent form (Appendix A) were analyzed in this study. A $5 gift card to a major national retail store was offered as an incentive for students to return a completed informed-consent form, irrespective of whether they agreed to participate in the study or not. Teachers in participating classrooms were also provided with a small stipend for classroom supplies as a reward for returning the informed-consent forms. These incentives were used to motivate teachers to encourage their students to put forth their best efforts when completing the assessment.

Data Collection

The participants in the study completed the LOCUS assessments in their respective schools during one 90-minute session or two 45-minute sessions. The completed exam booklets were sent back to the University of Florida and were stored in a locked room. CR item answers were recorded by the participants directly in the exam booklet using pen or pencil. The MC items were recorded by the participants using a bubble sheet that was scored using a machine. All MC responses and scores were recorded and stored digitally.

Responses to CR items were qualitative data in their original form, but were converted to quantitative data using the scoring process described in the next section. Once scores for the CR items were completed, they were stored with the corresponding MC scores from the same test booklet and all participant identifying information was removed from the dataset. Scoring information for the CR items was used to explore both research questions. In addition to scoring information, demographic data was collected for each participant and included their grade level, self-reported gender, race, ethnicity, and primary language used in their household. Demographic

information was used to control for variability when exploring the second research question that examined the relationship between overall conceptual understanding of statistics and understanding of variability.

Data Analysis

All parts of LOCUS CR items used in this study were coded by the element and perspective of variability that they addressed. Codes were determined using the descriptors from the Framework for Robust Understanding of Statistical Variation (Peters, 2011). As part of the LOCUS project, complete descriptions of items were made available online ("Professional Development," n.d.) that included the full item, an overview of the question, CCSSM standards addressed, ideal responses and scoring instructions, sample responses that indicated solid understanding, and common misunderstandings. These pieces of information were used to determine exactly what each part of an item was intended to address. Once the online item information was thoroughly considered, the descriptors from the Framework for Robust Understanding of Statistical Variation were used to match the item with the element and perspective of variability it best assessed.

An item that addressed variability from the design perspective was coded in the form DPx and one that addressed variability from the data-centric perspective was coded in the form DCPx. The x at the end of the code represented the element of variability that the item addressed. A 1 represented items that considered variational disposition, 2 represented items that considered variability in data for contextual variables, 3 represented items that addressed variability and relationships among data and variables, and 4 represented items that dealt with the effects of sample size on variability. For example, an item that considered variational disposition from the data-centric perspective was coded DCP1.

A second coder was briefed on the general process used to code the items. They utilized the LOCUS website and the Robust Understanding of Statistical Variation framework descriptors to independently code the items. After the first round of coding, there was 73% agreement among the 15 item parts. For all parts (11/15) with initial codes from the framework, the two coders agreed. For the other four, one coder assigned a code to the part while the other coder did not feel the part required understanding of variability to respond to. After further discussion, the two coders came to complete agreement for all item part codes. Detailed examples of how the items were coded and instructions for scoring each item will be provided in the next sections.

Item Coding

The first constructed response item on the form was referred to as the department store problem (Figure 3-1). The question required students to recognize the need for random selection when taking a sample and, recognize and describe why random assignment is necessary when conducting an experiment ("Professional Development," n.d.). Since the item deals with decisions involved with designing a study, it most closely aligns with the design perspective. To determine which element within the design perspective the item assesses, the ideal responses and scoring guidelines were used in conjunction with the question overview and sample student responses. The ideal response indicated that a method using random selection should be used to determine which credit card holders should be used in the sample. Controlling variability when designing a study using random selection appeared in the element labeled variability and relationships among data and variables in the framework. Therefore, the first part of the item was coded as DP3. The second part of the item required students to describe why random assignment to treatments is important in an experimental design. This part of the item also addressed DP3 because it involved controlling variability through random assignment in the study design.

1. A department store manager wants to know which of two advertisements is more effective in increasing sales among people who have a credit card with the store. A sample of 100 people will be selected from the 5,300 people who have a credit card with the store. Each person in the sample will be called and read one of the two advertisements. It will then be determined if the credit card holder makes a purchase at the department store within two weeks of receiving the call.

    a) Describe the method you would use to determine which credit card holders should be included in the sample. Provide enough detail so that someone else would be able to carry out your method.

    b) For each person in the sample, the department store manager will flip a coin. If it lands heads up, advertisement A will be read. If it lands tails up, advertisement B will be read. Why would the manager use this method to decide which advertisement is read to each person?

Figure 3-1. The department store problem.

The second constructed response item was referred to as the student council problem and took students through a short statistical investigation by having them write a survey question, display the data graphically, and draw a conclusion from the data. The first part of the item required students to write a survey question. A survey question in this scenario is used to answer a statistical question, which should anticipate variability. While the anticipation of variability is not directly necessary to answer the item, it is a crucial piece of understanding to be able to adequately create a survey question. Therefore, part (a) of the item was coded as DP1 because it required an acknowledgement of variability in pieces of the study design. Part (b) required students to provide a method they would use to sample 100 students to answer the survey question they wrote in part (a). Like the department store problem, this part of the item required students to recognize the need for random sampling and was therefore coded as DP3. Part (c) asked students to create a reasonable graphical display of possible responses to the survey. Creating a reasonable hypothetical dataset required a variational disposition from the data-centric

perspective because graphical displays should show reasonable variability in the data according to the context. Therefore, this part of the item was coded DCP1. Part (d) required students to recommend an activity and justify their answer using their graphical display in part (c), which involved choosing the activity with the most votes. This part of the item did not ask respondents to consider variability in their answer and was not coded for this study.

2. The student council members at a large middle school have been asked to recommend an activity to be added to physical education classes next year. They decide to survey 100 students and ask them to choose their favorite among the following activities: kickball, tennis, yoga, or dance.

    a) What question should be asked on the survey? Write the question as it would appear on the survey.

    b) Describe the process you would use to select a sample of 100 students to answer your question.

    c) Create a table or graph summarizing possible responses from the survey. The table or graph should be reasonable for this situation.

    d) What activity should the student council recommend be added to physical education classes next year? Justify your choice based on your answer to part (c).

Figure 3-2. The student council problem.

The third item, the boss preference problem, assessed the ability to read and understand a two-way table to see if there was an association between gender and preference for a male or female boss when starting a new job. The three parts of this item did not require students to show their understanding of variation. Parts (a) and (b) of this item asked students to read information directly from the table and describe the association between the two variables, respectively. These parts did not require the student to understand variation and were not coded. Part (c) asked students to look at a 2-way table from a different city to determine which data shows a stronger

association between the two variables. The way the question was presented in this item did not require understanding of variation and was not coded.

3. A researcher wanted to know whether there is an association between gender and preference for a male or female boss when a person starts a new job. The researcher randomly sampled 1,000 employed adults living in City A. Each person was asked, "If you were to start a new job, would you prefer to have a male boss or a female boss?" The table below shows the results for City A.

City A

	Prefer Male Boss	Prefer Female Boss	Total
Male	428	122	550
Female	272	178	450
Total	700	300	1000

a) i) What percentage of males prefer a male boss?
(ii) What percentage of females prefer a male boss?

b) Describe the association between gender (male or female) and preference for a male or female boss in City A.

The researcher asked the same question in a survey of 1000 randomly selected adults from City B. The table below shows the results for City B.

City B

	Prefer Male Boss	Prefer Female Boss	Total
Male	425	175	600
Female	275	125	400
Total	700	300	1000

c) In which city, A or B, is there a stronger association between gender (male or female) and preference for a male or female boss? Justify your response.

Figure 3-3. The boss preference problem.

4. A study was carried out to investigate whether there is a relationship between the percent of hearing loss and the volume at which people typically listen to music. Ten high school students agreed to participate in a study. Each was given a music player with headphones and was asked to listen to music for 10 minutes. The students were told to adjust the volume to a comfortable setting. After 10 minutes, the volume setting, which ranges from 1 to 10, was observed for each student. Each student then took a hearing test, and a measure of hearing loss (in percent) was recorded. The data are shown in the table below.

Volume Setting (x)	8	10	1	4	5	8	3	1	2	8
Hearing Loss (y)	23	24	11	9	15	19	14	5	7	15

   a) Construct an appropriate graphical display that allows you to investigate the relationship between volume setting and hearing loss.

   b) Based on the graphical display, describe the relationship between volume setting and hearing loss.

   c) From this study, is it reasonable to conclude that listening to music at a high-volume causes hearing loss? Explain why or why not.

Figure 3-4. The hearing loss problem.

Item four, the hearing loss problem, presented data from an experiment to determine the relationship between hearing loss and the volume at which people listen to music. Part (a) had students use the given dataset to construct a graphical display that would allow them to investigate the relationship between the two variables. This part of the item did not require understanding of variation to complete and was not coded. Part (b) asked students to use their graphical display to describe the relationship between hearing loss and volume. An adequate description of the relationship between the two variables may require understanding of variation from the modeling perspective, which was not the focus of this study, and therefore this part was not coded. Part (c) required students to conclude whether listening to music at high volume

causes hearing loss. To answer this part, students must refer to the study design introduced in the item stem to notice this is an observational study where cause-and-effect conclusions cannot be made. Considering study design when making conclusions involved understanding of statistical variation from the design perspective. This item was coded as DP3 because it involved critiquing the extent to which variability was controlled through the study design. In this case, the observational study design did not control for confounding variables that may have interfered with the relationship between volume and hearing loss, and therefore a cause-and-effect conclusion between the two variables could be drawn.

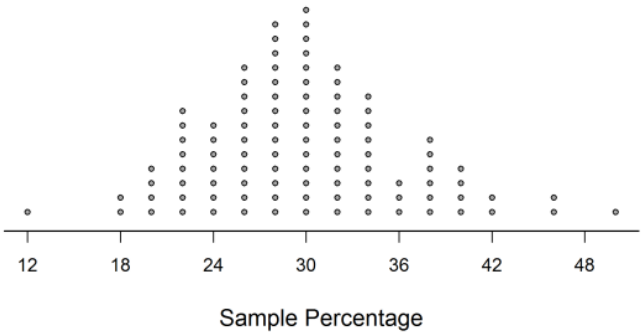
The final item on the constructed response, the school day problem, followed an investigation to test the claim by a national newspaper that 30% of students favor an extended school day. Part (a) had students complete the necessary details for conducting a simulation to see what sample percentages would be expected if the population percentage was actually 30%. This did not require understanding of variation from the design or data-centric perspective and was not coded. Part (b) involved the use of the given sampling distribution of the sample proportion to determine if an observed result was plausible. To determine plausibility, the student had to consider the variability in the sampling distribution. Calculating the probability that a result as or more extreme than the observed result required an understanding of how the data vary. Therefore, this part was coded DCP3 because it explored the variability in a visual representation of the data. Part (c) had students draw a conclusion based on the sample data in the problem. The justification for the conclusion could be based on the sampling variability, and would therefore examine variability and relationships among data and variables from the data-centric perspective. This part of the item was coded as DCP3.

5. Stella saw the following headline in a national newspaper: "30 Percent of High School Students Favor Extended School Day." She wondered if the percentage of students at her school who favor an extended school day was less than 30 percent. To investigate, she selected a random sample of 50 students from the 1,200 students at her school and asked each student in the sample if he or she favors an extended school day.

   Only 12 of the students in the sample favored an extended school day. Because the sample percentage is (12/50)100 = 24%, Stella thinks that fewer than 30 percent of the students at her school favor an extended school day. She wonders if it would be surprising to see a sample percentage of 24 or less if the school percentage is really 30.

   a) To see what values of the sample percentage would be expected if the school percentage was 30, she decides to use 1,200 beads to represent a student who favors an extended school day and a white bead to represent a student who does not. How many red beads and how many white beads should Stella use?

      Stella put all the beads in a box. After mixing the beads, she selected 50 of them and computed the percentage of red beads. She put the 50 beads back in the box and repeated this process 99 more times. Then, she made the following dotplot of the 100 sample percentages:



   b) If the school percentage were actually 30%, how surprising would it be to see a sample percentage of 24% or less? Justify your answer using the dotplot.

   c) Based on her sample data, should Stella conclude that the percentage of students at the school who favor an extended school day is less than 30%? Explain why or why not.

Figure 3-5. The school day problem.

The final item codes resulted in eight item parts that addressed 4 cells from the robust understanding of statistical variation framework. The only two elements that were covered by these item parts were variational disposition and variability and relationships among data and

variables. Only one item part addressed each of the DP1 and DCP1 cells of the framework, 4 item parts addressed the DP3 cell, and 2 item parts addressed the DCP3 cell. The parts of the framework considered in this study are shown in Table 3-2.

Table 3-3. Components of framework relevant to this study.

	Design Perspective	Data-Centric Perspective
Variational disposition	DP1: Acknowledging the existence of variability and the need for study design in d. controlling the effects of variation from extraneous variable(s); a. including considerations of variation for variable(s) of interest during data analysis; or b. using sample statistics to infer population parameters for the variable(s) of interest	DCP1: Anticipating reasonable variability in data by d. considering the context of data; e. recognizing that data descriptions should include descriptions or measures of variability (and center); or f. recognizing unreasonable variability in data (e.g., that which could result from a data entry error)
Variability and relationships among data and variables	DP3: Controlling variability when (1) designing studies or (2) critiquing the extent to which variability was controlled in studies by c. using random assignment or random selection of experimental/ observational units to (in theory) equally distribute the effects of uncontrollable or unidentified sources of variability; or d. using study design to control the effects of extraneous variables (e.g., by incorporating blocking in experimental design or stratifying in sampling designs) to isolate the characteristics of the variable(s) of interest or to isolate systematic variation from random variation	DCP3: Exploring controlled and random variability to infer relationships among data and variables by e. (1) using and (2) interpreting patterns of variability in various representations of data; f. focusing on aggregate or holistic features of variability in data to make comparisons; g. (1) using or (2) interpreting appropriate summary measures of the variability in data to make comparisons (e.g., transformed versus untransformed data); or h. examining the variability within and among groups

Item Scoring

A scoring procedure for the CR items was developed as a minor extension of the item codes. For each part of an item, if the student response showed evidence of understanding of variability according to how the part was coded, the part received a score of 1. Otherwise, the

part received a score of 0. In some cases, responses could receive a score of 0.5 if they displayed a developing understanding from the relevant cell. The only example of a response that received a score of 0.5 was for cell DP3 when stratified sampling was discussed, but no randomness was involved in the description. Evidence of understanding was determined by the presence of key features in the response that addressed variability according to the framework and did not factor in the quality or depth of the response. For example, a response that suggested the use of a random sample for part (a) of the department store problem and a response that describes the complete process required to take a random sample would both receive a score of 1. Prior research on responses to LOCUS items suggested that high-quality and diverse responses to CR items were rare (Case & Jacobbe, 2014; Foti & Jacobbe, 2015; Whitaker & Jacobbe, 2014; Jacobbe et al., 2015).

The ideal and sample student responses from the LOCUS website were used as guidelines to develop examples for each item part that would indicate understanding according to the item code. For example, the scoring procedure for the department store problem was as follows:

Item 1: Department Store:

> Part (a) - Score a 1 for DP3 if response addresses controlling for variability in the study design using random selection. Score a 0.5 if the response addresses an appropriate stratification technique, but does not address randomness. Otherwise, score a 0.
>
> e.g., Random selection is used to select the sample, and a method is described for how this could be done by numbering the credit card holders from 1 to 5300; Sample is stratified into frequent and non-frequent shoppers before random sampling is conducted.
>
> Part (b) - Score a 1 for DP3 if response explains why random assignment to treatments is important in the design of a statistical experiment. Otherwise, score a 0.

e.g., indicates that flipping a coin to determine which advertisement is read results in random assignment to treatments and therefore the study can conclude cause and effect

An example student response from the LOCUS website is shown in Figure 3-6. For part (a), the response adequately described a process for random selection and would be scored a 1 for code DP3. In part (b), the response indicated that random assignment helps to control variability and results in a conclusion that could establish a cause and effect relationship. Thus, this part would be scored a 1 for code DP3.

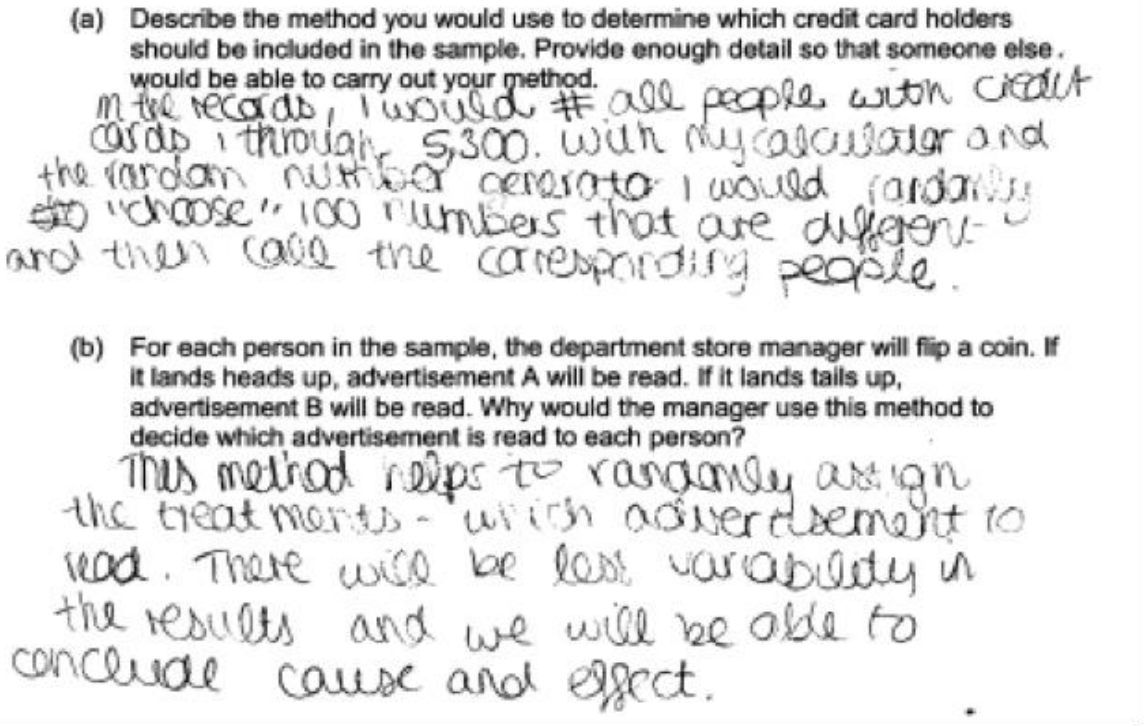


Figure 3-6. Sample student response from LOCUS website.

During the scoring of the items, notes were taken on both common and uncommon responses to improve consistency of scores and to allow for responses to show evidence of understanding outside of the cell the item was coded for. On extremely rare occasions, a response would show evidence of understanding of variability from a cell of the framework that

differed from the cell the item part was coded for. These responses were recorded, however, none of them occurred with a frequency high enough to yield meaningful interpretations in this study. Each time a unique decision was made about the evidence of understanding shown in a response, the booklet number, quoted response, and subsequent decision was recorded. These notes were used to update the full scoring procedure to be more descriptive in terms of observable features of a response. The full scoring procedure, updated with information from the scoring, is shown in in Appendix B.

Table 3-4. Sample of scoring table for a participant (not actual data).

	Design Perspective	Data-Centric Perspective	Totals
variational disposition	DP1: 0/1 - 0%	DCP1: 1/1 - 100%	1/2 - 50%
variability in data for contextual variables	-	-	-
variability and relationships among data and variables	DP3: 2/4 - 50%	DCP3: 2/2 - 100%	4/6 - 67%
effects of sample size on variability	-	-	-
Totals	2/5 - 40%	3/3 - 100%	5/8 - 63%

Scoring of CR items was done by the author of this dissertation as well as a second independent coder that was trained to use the procedure. Inter-rater agreement was assessed using agreement percentages and is discussed in Chapter 4. Once all responses were scored, each participant had a 2x4 matrix representing their scores in each of the two perspectives (columns) and four elements of variation (rows). A hypothetical example of a scored exam matrix can be seen in Table 3-4.

Research Question 1

The first research question sought to provide a snapshot of how U.S. high school students understand the concept of variability. The results from the scoring procedure were aggregated across students and converted to proportions. Proportions in each individual cell of the framework represented an estimate of how well high school students understood an element of variability from a particular perspective, row sums represented how well students understood an element of variability across perspectives, and column sums represented how well students understood variability from a particular perspective across all elements. Aggregating across all rows and columns resulted in a single estimate for how well students generally understood variability. Direct interpretations of a proportion would be the proportion of responses that showed evidence of understanding from a part of the framework (e.g., cell, perspective, element).

Part of the snapshot involved testing whether one perspective seemed to prompt more evidence of understanding than the other. A statistical significance test was used to test for differences in understanding in the design and data-centric perspectives. A difference between two-proportions *z*-test was utilized under the null hypothesis of no differences to see if there were significant differences between the variables for high school students. Additionally, two item parts that were nearly identical, parts 1a and 2b, were analyzed in depth because of differences observed during scoring. A difference between two-proportions *z*-test was used under the null hypothesis of no difference to determine if responses were showing evidence of understanding at different frequencies despite being nearly identical items.

Research Question 2

The second research question sought to determine if there was a relationship between overall understanding of statistics and understanding of variability, and if so, the strength of that

relationship. A multiple linear regression model of the form $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$ was fit to test this relationship, where y was the observed overall understanding of statistics, x_1 through x_k were the observed variables for k demographic features, x_p was the observed understanding of variability, β_0 through β_p were the model parameters to be estimated, and ϵ was the random error. Model fitting was conducted in the R statistical software package and followed standard procedures for multiple linear regression (Rawlings, Pantula, & Dickey, 2001).

Overall understanding of statistics was measured by the score students received on the MC portion of LOCUS. Only the MC items were considered in their LOCUS score to avoid correlations between correctness and evidence of understanding of variability in the CR items. Understanding of variability was measured by the score students receive on the CR items according to the scoring procedure used in this study, which was different than the original scoring rubric used as part of the LOCUS project. Additionally, the model included the demographic information that was collected on the students such as grade, gender, race, ethnicity, and primary language spoken at home. While these demographics were not the primary focus of this study, significant effects needed to be controlled for to limit the influence of outside variables on the relationship between understanding of statistics and variability.

Two models were considered to attempt to better understand the relationship between understanding of variability and overall understanding of statistics. The initial model aggregated scores in both the design and data-centric perspectives and used a single score for understanding of variability. Another model split understanding of variability into two variables, one for understanding from the design perspective and one for understanding from the data-centric perspective. The purpose of this model was to see if a particular perspective had a stronger

impact on overall understanding of statistics than the other. The adjusted R^2 value was used to determine how much of the variability in the MC scores could be attributed to the regression on the predictors.

Limitations

Despite careful consideration throughout this study, there were certain limitations that must be discussed to properly interpret the potential findings. To start, this study used a carefully planned quantitative methodology, however, there were details regarding the sampling methods that must be addressed. The sample of students was not chosen randomly from the population of grades 9-12 students in the United States that have taken a course involving statistics. Therefore, it is not technically permissible to generalize the results of this study beyond the participants involved. The results of the data analysis could only suggest conclusions about the population. With data collected from over 700 students, the conclusions drawn from this study were informative, but further studies must be conducted if the patterns seen are to be confirmed.

A truly random sample of United States secondary students could potentially include many students that have not had any formal statistics education in their school career. This would, most likely, bias the results towards a lack of understanding of variability. On the other hand, because the schools and classrooms that were chosen for this study were selected because of their connection to researchers in statistics education and were from high performing districts, the results of this study may just as easily overestimate the level of understanding of variability in secondary students. Thus, the findings in this study must be interpreted with caution and be taken as a piece of evidence toward more general conclusions.

This study considered the use of post-stratification sample weights (Dillman et al., 2014) to allow for unbiased generalizations to a larger population. Weighting the sample to the population of U.S high school students seemed to be too large of a deviation from the original

sampling methods. The most logical choice of population would be U.S. high school students that have taken a course involving statistics, however, demographic data for this population was not readily available. Such data exists for students that took AP statistics, however, there is no reason to believe that the AP statistics population looks anything like the population of students that have taken any course involving statistics. Thus, methods to weight the sample were avoided in this study and the existing limitations on generalizations remained.

Another limitation to this study was that pilot administrations of the LOCUS assessments revealed some potential issues regarding the difficulty of the items. Many students in the pilot administration of LOCUS received very low scores, particularly on the CR items (see Case & Jacobbe, 2014; Foti & Jacobbe, 2015; Whitaker & Jacobbe, 2014; Jacobbe et al., 2015). One possibility is that the LOCUS items were at a level of understanding that was too far above the understanding of most students in the current secondary environment. Another possibility was that due to the focus of the LOCUS assessments on conceptual understanding, many of the items were lengthy word problems. This study did not consider the possibility that students may not have had the proper language proficiency to accurately display their understanding of variation. Thus, it is possible that the results of this study regarding students' understanding of variability are confounded with language ability.

Finally, one of the purposes of this study was to provide an example of using assessment items as tasks to analyze students' understanding of variation. However, because the CR items are not entirely open ended, and students were not given the opportunity to more thoroughly explain their answers (e.g., in an interview), their ability to fully display their understanding of variability may have been limited. Therefore, the results of this study should be used as evidence of potential, versus definite, weaknesses and strengths in current secondary statistics education

efforts. This limitation was reflected in the data analysis and interpretation of the results by examining responses for evidence of understanding instead of strictly thinking of responses as correct or incorrect.

CHAPTER 4
RESULTS

In this chapter, the results of the data analysis will be presented. Data were collected and analyzed in alignment with the goals of this study. The fundamental goal of the data analysis was to learn more about how U.S. high school students understand the concept of variability, as defined by the Framework for Robust Understanding of Variation (Peters, 2011). The data consisted of scored LOCUS CR items according to the scoring procedure outlined in Chapter 3. Contrary to common item scoring, where a 1 represents a correct response and a 0 represents an incorrect response, the scores for the items represent whether a response showed evidence of understanding of variability. Therefore, a response that received a 0 simply means that it did not provide sufficient evidence of understanding.

The following subsections of this chapter display the results of the data analysis as they pertain to the research questions guiding this study. Results from an inter-rater agreement process are also presented. Additionally, the nature of missing data from the collection phase and the techniques used to rectify the issues of missing data are discussed. Results of the data analysis are presented with interpretation in this chapter, and their implications will be discussed further in Chapter 5 of this dissertation.

Inter-Rater Agreement

To provide evidence for consistency in the scoring procedure, a second scoring was completed by another graduate student in statistics education. A meeting was held to thoroughly discuss the Framework for Robust Understanding of Variation (Peters, 2011), each CR item used in the study, and the scoring procedure presented in the previous chapter. Sample test booklets were examined to practice using the scoring procedure until the second scorer felt comfortable with the complete procedure. A random sample of 78, equal to approximately 10% of the total

number of test booklets, was then provided to the second scorer to be reviewed using the scoring procedure.

A comparison of the two scorers' data revealed that out of 624 total scored item parts, the two scores were in exact agreement 93.4% of the time. Since some item parts were scored as 0.5 when partial evidence of understanding of variability was presented, the analysis of scoring agreement also considered agreement within 0.5 points. Under this condition, 96.6% of the scores were in agreement. Percentage of agreement has received criticism as an index for inter-rater agreement because it does not consider agreement that would happen by chance (Cohen, 1960; Robinson, 1957).

Cohen's (1960) kappa takes agreement that would occur by random chance into account and is defined by

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement. For the two scorers in this study, kappa was calculated to be 0.88, meaning the two scorers were in agreement approximately 88% of the time when accounting for agreements due to random chance.

Missing Data

The scored data set contains missing data for CR items and, less commonly, for demographic data. Enders (2010) describes how partial deletion, removing data until there is no missing information left, and single imputation, techniques that estimate the missing data in one instance, yield suboptimal results. Multiple imputation, which involves techniques to estimate the missing data more than once, is the recommended method for dealing with missing data (Enders, 2010; van Buuren & Groothuis-Oudshoorn, 2011).

The missing data pattern, shown in Table 4-1, reveals the number of complete cases in the dataset and the number of cases that have various amounts of missing data. Typically, in order for an imputed dataset to yield unbiased results, the data needs to be missing completely at random (MCAR) and must not contain large amounts of missing data. To test for MCAR, Little's test (1988) was used and resulted in a p-value of approximately 0. While failing to reject the null hypothesis of MCAR provides evidence that the data is MCAR, it does not prove that it is. Students with lower understanding of statistics or those who put less effort into LOCUS were more likely to have missing data. Therefore, it did not necessarily make sense for this data to be MCAR. Additionally, no item part had greater than 10% missing data, with most having under 6% of their data missing.

Predictive mean matching in the MICE package in R (van Buuren & Groothuis-Oudshoorn, 2011) was used to multiply impute the missing data. Reviewing the summary statistics for the imputed datasets and a dataset of only the complete cases showed similar results. In total, five imputed data sets were constructed for use in multiple linear regression models. Each imputed data set was used to create a linear regression model and the coefficients from the multiple models were pooled to find the final coefficient estimates.

Table 4-1. Missing data pattern from full dataset.

# of Data Missing	0	1	2	4	5 or more
# of Cases	492	26	2	20	85

Descriptive Statistics

Once the imputed datasets were created, descriptive statistics were computed for the understanding of variability on items and perspectives, and are shown in Table 4-2. Item parts are organized into their perspective and element from the robust understanding of statistical variation framework. The mean of each item part across all students, which is equivalent to the proportion of students that showed evidence of understanding of variability for that part, is presented. Item level results will be presented in more detail in the following sections.

Table 4-2. Descriptive statistics for all item parts.

	Design Perspective					Data-Centric Perspective		
	DP1	DP3				DCP1	DCP3	
	2a	1a	1b	2b	4c	2c	5b	5c
Mean/Proportion showing evidence of understanding	.94	.62	.04	.75	.14	.93	.31	.07
Showed evidence of understanding	593	391	25	473	88	587	196	44
Did not show evidence of understanding	38	240	606	158	543	44	435	587

The correlation matrix for the 8 item parts are shown in Table 4-3 and are also grouped by the perspective they address. Item part-to-item part correlations were generally very low, which provided evidence that each item part presented different information in terms of understanding of variability. More likely, the correlations were low because of the nature of the scores on the responses. For example, nearly all students showed evidence of understanding of variability on items 2a and 2c, while virtually no students showed evidence on items 1b, 4c, and 5c. Item parts with polarized scores had no room to provide evidence of correlation between the item parts. The item parts where a decent number of students both did and did not show evidence

of understanding of variability were 1a, 2b, and 5b. 1a and 2b had the highest correlation of all the items, which may be due to them coming from the same perspective or, more likely, because they were nearly identical item parts. Further results regarding the relationship between these two items are discussed later in this chapter. Correlations between 1a and 5b, and 2b and 5b were also low, however, this was expected because the items addressed different perspectives of variability.

Table 4-3. Item part-to-item part correlations for all items, organized according to perspective.

Correlation Matrix for all LOCUS CR Items Organized by Perspective			Design Perspective					Data-Centric Perspective		
			DP1	DP3				DCP1	DCP3	
			2a	1a	1b	2b	4c	2c	5b	5c
Design Perspective	DP1	2a	-							
	DP3	1a	.064	-						
		1b	.050	.062	-					
		2b	.121	.350	.092	-				
		4c	.084	.224	.227	.150	-			
Data-Centric Perspective	DCP1	2c	.188	.126	.023	.168	.061	-		
	DCP3	5b	.127	.234	.152	.159	.214	.135	-	
		5c	.069	.204	.206	.107	.315	.028	.286	-

The LOCUS MC scores, which were scored independently of this study, can be seen in the histogram in Figure 4-1. These scores were used as a measurement of overall understanding of statistics, and the maximum possible score was 23. The mean score on the MC was 11.96 with a standard deviation of 4.46, and the dataset appears to be slightly skewed right.

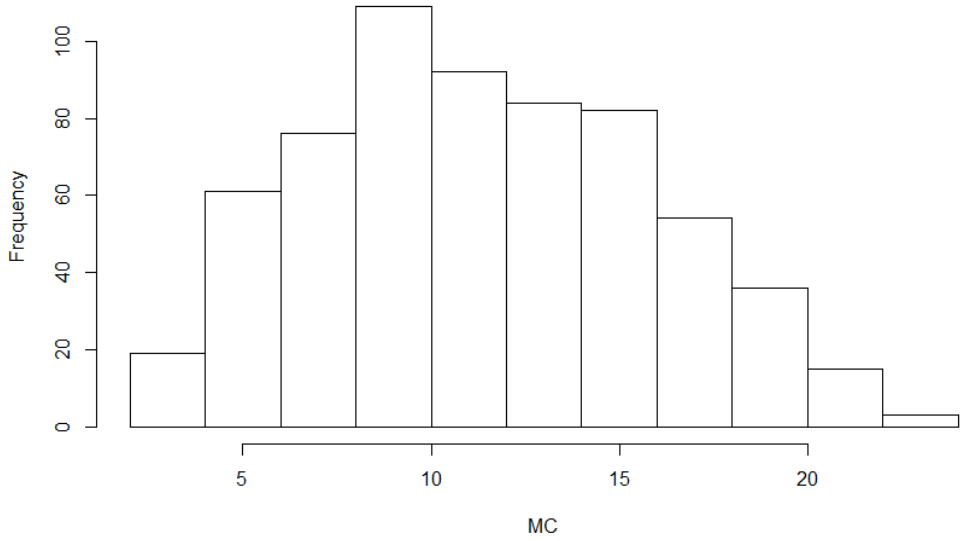


Figure 4-1. Histogram of the multiple-choice scores from LOCUS (scored independently from this study).

Understanding of Variability

In total, 742 high school students in grades 9, 10, 11, or 12 participated in the Intermediate/Advanced Form 2 administration of the LOCUS assessment. Of those participants, 631 were included in the data analysis because they gave consent and answered at least 1 CR item. Understanding of variability was considered from various degrees of resolution. The data was analyzed at the perspective, element, and item levels. The results of these analyses are presented in this section and a summary of results are shown in Table 4-4.

Table 4-4. Proportion of item parts that showed understanding in Robust Understanding of Variation Framework.

Element/Perspective	Design Perspective	Data-Centric Perspective	Totals
Variational disposition	.940	.928	.934
Variability and relationships among data and variables	.389	.191	.323
Totals	.499	.437	.476

Summary scores across rows and columns are not equally weighted across all cells of the framework. For example, of the 50% of responses from the design perspective that showed evidence of understanding of variability, nearly all of them (4/5) came from the variability and relationships among data and variables element. Reporting the scores in this manner implies that more weight was given to elements with more items. This study placed less weight on variational disposition than variability and relationships among data and variables. The stronger emphasis on variability and relationships among data and variables is supported by the deeper understanding of variability required in that element. Variational disposition focuses on acknowledging and anticipating variability, which require lower-levels of understanding than considering, describing, and measuring variability in data. Additionally, the LOCUS assessments development process using ECD resulted in more items that addressed the variability and relationships among data and variables element, suggesting a stronger emphasis on this element is warranted.

At the perspective level, the data showed more evidence of understanding from the design perspective than the data-centric perspective. Aggregating scores within the perspectives revealed that nearly 50% of all possible responses provided evidence of understanding from the design perspective, compared to 44% of responses from the data-centric perspective. A hypothesis test was conducted to test for a significant difference in understanding from one perspective. Under the null hypothesis of no difference between the two perspectives, a difference between two proportions test was conducted. The 95% confidence interval for the difference between the design perspective and data-centric perspective was (0.034, 0.091). Students showed significantly more evidence of understanding from the design than the data-centric perspective.

The design perspective consisted of 5 total item parts and the data-centric perspective consisted of 3 total item parts. Therefore, when comparing proportions, it was important to recognize that the perspective-level scores achieved by individuals were discrete in nature. For example, in the data-centric perspective, individuals could only achieve 0, 33%, 66%, or 100% of the possible points because there were only 3 parts that addressed the perspective. Despite this realization, the difference between the two perspectives was still quite high for this large sample. A breakdown of points scored for each of the two perspectives can be seen in Figure 4-2.

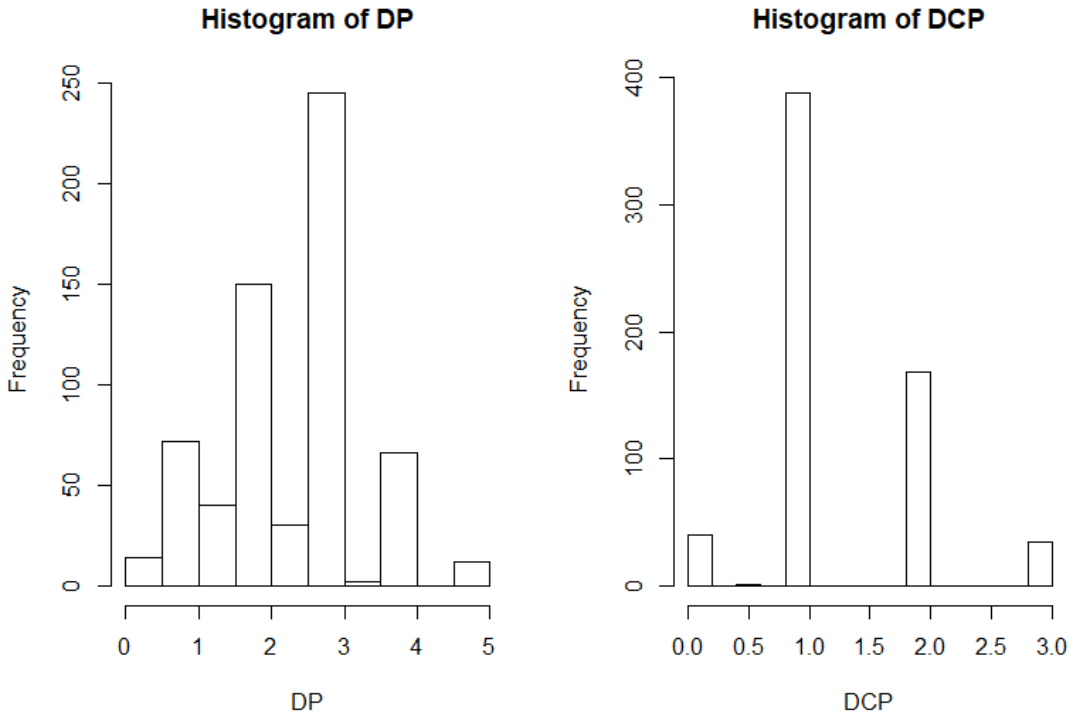


Figure 4-2. Histograms of scores on the DP (out of 5) and DCP (out of 3). Note: vertical axes are on different scales.

In the design perspective, a majority (67.4%) of students fell right in the middle and showed evidence of understanding of variability on between 2 and 3 out of 5 possible item parts. In the context of this study, this means that students had a middle-of-the-road understanding of variability as it pertained to the design of studies. Specifically, the elements of the design

perspective on LOCUS addressed acknowledging the existence of variability and knowing how to control some of it through study design. These are pivotal parts of the statistical problem-solving process according to the formulating questions and collecting data process components of the K-12 GAISE framework (Franklin et al., 2007).

In the data-centric perspective, nearly 61% of students scored a 1 out of 3 possible points. Students did not show strong evidence of understanding from this perspective, however, two of the three parts were on an item that carried out a simulation-based hypothesis test. It is entirely possible that the simulation was an unfamiliar scenario that distracted them from showing evidence of understanding of variability. Through item part 2c, students did show very strong evidence of their ability to anticipate reasonable variability in data, the descriptor for DCP1. Just over 5% of students showed evidence of understanding on all 3 parts from the data-centric perspective. These students showed evidence of having a Level C understanding of analyzing and interpreting data, which placed them in the highest level of understanding in the GAISE framework (Franklin et al., 2007).

The variational disposition element of the framework was assessed through two item parts: one from the design perspective and one from the data-centric perspective. Approximately 93.4% of all possible responses showed evidence of understanding from this element of the framework. In other words, most students showed their ability to acknowledge and anticipate variability. The second element that was assessed through the CR items was variability and relationships among data and variables. Across the 6 item parts that addressed this element of variability, approximately 32.3% of responses showed evidence of understanding. Item parts that addressed this element saw less evidence of understanding across both perspectives, however, responses showed more evidence of understanding, in this element, from the design perspective

(0.39) than from the data-centric perspective (0.19). The differences in evidence of understanding between design and data-centric perspectives mostly appeared in this element of variability. In the design perspective, understanding of variability and relationships among data and variables concerned the design of studies to control for variability through random selection of samples, random assignment of treatments in experiments, or designing studies to make causal conclusions. From the data-centric perspective, this element of variability involved exploring controlled and random variability to draw conclusions.

The four cells of the robust understanding of statistical variation framework addressed by the LOCUS CR items in this study were DP1, DP3, DCP1, and DCP3. The proportion of responses that showed understanding from these cells were about 0.94, 0.39, 0.93, and 0.19, respectively. The histograms in Figure 4-3 display the breakdown of the total scores on each of the four cells of the framework.

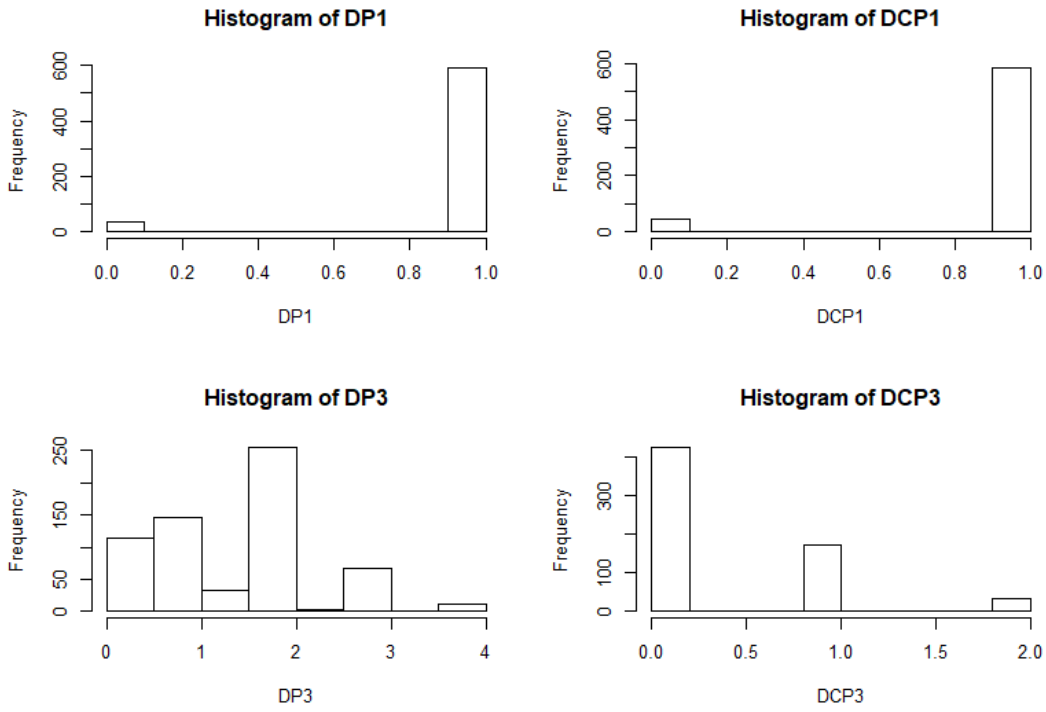


Figure 4-3. Histograms of scores for each cell of the framework.

DP1 was only relevant to item 2a, and nearly all of students' responses (94%) showed evidence of understanding. Specifically, students showed evidence of anticipation of variability when using a survey question to help answer a statistical question. While this result suggests that nearly all students could present a relevant survey question given a scenario, the item part did not directly address their ability to formulate a statistical question. This one item part may therefore overestimate students' understanding of DP1. Empirical research on students' ability to formulate statistical questions is scarce. In one study, Ben-Zvi (2002) found that nearly 60% of forty 7th grade students could provide research questions that focused on overall patterns in data when given a contextual scenario.

DCP1 was assessed through a single item part and had a similarly high proportion of students show evidence of understanding (93%). This item part required students to create a display of hypothetical data that might result from the survey question used in the previous parts. Students showed evidence of understanding from DCP1 by displaying a set of data where survey responses varied. While reservation should be used when concluding understanding of DCP1 from one item part, the item more directly addressed DCP1 than in the case of DP1. Shaughnessy et al. (1999) and Reading and Shaughnessy (2000) both discovered that students had trouble anticipating variability in sampling situations. While the scenario in the LOCUS item was less complex, nearly all students could anticipate variability in the hypothetical outcome of sampling.

Evidence of understanding of variability from DP3 was discovered in four different item parts: 1a, 1b, 2b, and 4c. About 87% of students scored between 0 and 2 out of the 4 possible points for DP3. On 1a, about 62% of responses showed evidence of understanding from DP3 by anticipating variability in a study design and knowing to implement randomness into their sampling technique. Only about 4% of responses to 1b showed evidence of understanding of

variability from DP3 by describing that random assignment of treatments is important for controlling variability from outside sources. Item 2b, which was nearly identical to 1a, had about 75% of students show evidence of understanding of variability from DP3. On 4c, about 14% of students showed evidence of understanding from DP3 by stating that cause and effect conclusions require controlling for variability through an experimental design. These concepts of study design fall in both levels B and C of the GAISE framework (Franklin et al., 2007), meaning they are beyond the expected understanding of a novice statistics student. These concepts are vital to statistical literacy and the ability to interpret results from other studies (Gal, 2000).

Evidence of understanding of variability from DCP3 was discovered in two different item parts: 5b and 5c. On 5b, about 31% of students showed understanding of variability from DCP3 by acknowledging variability in a sampling distribution. Only 7% of students displayed evidence of understanding of DCP3 on item 5c by acknowledging that a difference between a given and an observed mean could be due to random variation. While more than one item part addressed DCP3, it is worth noting that the two item parts were from the same item. As noted above, this item's context was a simulation study that may be unfamiliar to many students. Additionally, the variability concepts in part 5b are only addressed in Level C of the K-12 GAISE framework (Franklin et al., 2007).

Links to Context

Because items 1a and 2b both asked for a process to use for collecting a sample from a population, they prompted further analysis. Aside from one item using the term "method" and the other using the term "process," the only difference between the two was the context of the item stem. Item 1 was looking for a sample of credit card holders from the 5,300 people that owned a credit card with a store. Item 2 was looking for a sample of students from a middle

school to conduct a survey. Research shows conflicting evidence regarding students' understanding of variability across contexts (Torok & Watson, 2000; Ben-Zvi, 2004; Reading, 2004).

In this instance, a hypothesis test was conducted to determine if the proportions of students that showed evidence of understanding of variability from DP3 in each of these two item parts were significantly different. The 95% confidence interval for the difference in population proportions was (-0.19, -0.08) and the p-value for the significance test was nearly 0. A statistically significantly higher proportion of responses showed evidence of understanding of DP3 on item 2b than on item 1a.

Torok & Watson (2000) found that no association existed between context and understanding of variability, as students displayed no difference in their understanding despite more familiarity with some of the tasks' contexts than others. However, this finding suggests there may be a link. Reading (2004) and Ben-Zvi (2004) have also showed evidence that context played a role in understanding of variability. In one case, the context intervened with students' abilities to use analysis techniques they had recently covered in class (Reading, 2004). In the other, context helped students connect their understanding of statistics to the problem (Ben-Zvi, 2004). It is unclear whether the context of the credit card in the LOCUS item distracted students, or the context of the middle school survey helped students focus on the sampling techniques.

Role of Variability in Understanding of Statistics

To address the second research question, a multiple linear regression model was fit. The dependent variable in the model was each student's LOCUS MC score. Models were fit with both DP score and DCP score as predictors. Demographic and grade level information was included in the model to control for any effects they may have had on the students' overall understanding of variability. The model that was fit to each of the 5 imputed datasets was $y_i =$

$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$, where y is the observed overall understanding of statistics, x_1 through x_k are the observed variables for k demographic features, x_p is the observed understanding of variability, β_0 through β_p are the model parameters to be estimated, and ϵ is the random error.

Overall scores for DP and DCP were used in the regression model. Since there were not an equal number of items representing each cell of the framework, some cells more heavily influenced the overall scores for the perspective. For example, in each students' DP score, DP3 items account for nearly 80% of the score. These weighted perspective scores were used because not all elements were determined to be equally important to understanding of variability and overall understanding of statistics. As described in the section regarding results for research question 1, these weighted perspective scores were justified through their representation in the LOCUS assessments.

Diagnostic plots for one of the imputed data sets are shown in Figure 4-4. The assumptions for fitting a multiple linear regression model visually appear to be satisfied in these plots. However, it may be worth noting the potential pattern in the residual plot, where residuals tend to be lower at the extremes of the dataset. Once the models were fit using each of the 5 imputed datasets, the coefficients were pooled to achieve the final estimates (Enders, 2010; van Buuren & Groothuis-Oudshoorn, 2011).

The estimates of the pooled model coefficients are shown below in Table 4-5. Controlling for all the demographic information and grade level, the coefficients for both DP and DCP were significantly greater than 0, with p-values < 0.001. For every 1 point increase in evidence of understanding of variability shown from the design perspective, students' MC scores increased by about 1.94 points, on average. For every 1 point increase in evidence of understanding of

variability from the data-centric perspective, students' MC scores increased by about 1.73 points, on average. With a maximum of 23 possible points, a mean of 11.96 and a standard deviation of 4.46 points, these predicted changes in MC scores equate to an increase of 0.43 and 0.39 standard deviations. The difference in predicted MC scores between a student that showed no understanding of variability from either perspective and a student that showed evidence of understanding on all item parts is nearly 15 points, on average. Confidence intervals for both perspectives showed that the average increase in MC score could be anywhere from just over 1 point to over 2 points per 1 point increase in evidence of understanding of variability from each perspective.

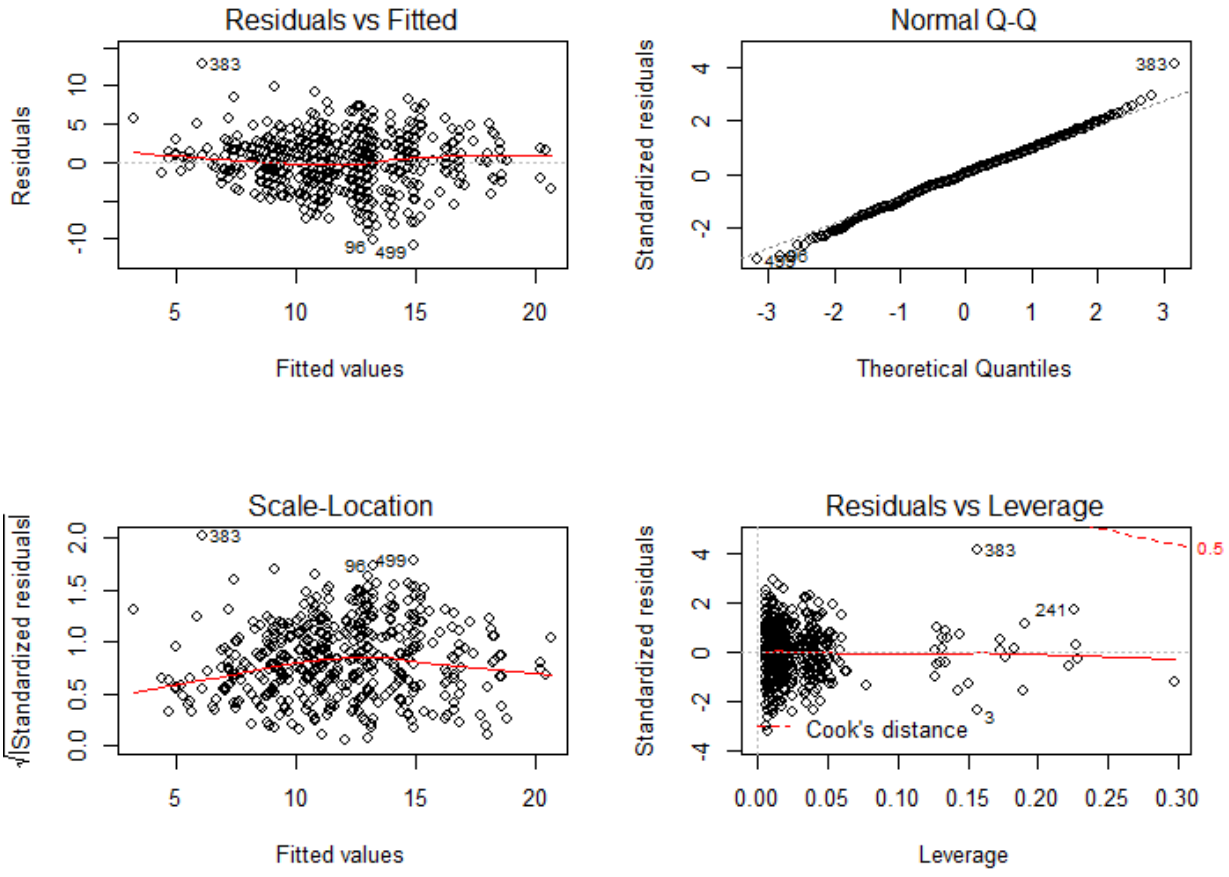


Figure 4-4. Diagnostic plots for the multiple linear regression model fit to one of the imputed datasets.

Table 4-5. Pooled estimates from multiple linear regression models on MC scores.

	Pooled Estimates		
	B	*CI*	*p*
DP	1.91	1.59 – 2.23	< 0.001
DCP	1.73	1.21 – 2.25	< 0.001
Observations		631	
R^2 / adj. R^2		.430 / .416	
Δ R^2	0.300	F = 161.478	p < 0.001

Given that the two perspectives have a statistically significant association with MC scores, fitting a model with overall understanding of variability as the aggregate of the two perspectives would yield no extra information. It is worth noting that the correlation between the two perspectives, overall, is 0.414. Thus, the two perspectives appear to provide different information to the estimated model. The coefficient of determination, R^2, for this model was only 0.430. In other words, only 43% of the variability in MC scores was explained by the variation in the linear combination of the predictors in the models. The adjusted R^2 value, which penalizes models for large number of explanatory variables, was 0.416.

A model comparison test was conducted to analyze the impact that adding DP and DCP scores as predictors had on the model. With a p-value of approximately 0, there was significant evidence that the model with the two predictors was significantly different than the reduced model that excluded them. The value of R^2 increased by 0.300 with the addition of the two predictors, providing further evidence that their addition to the model improves the predictive power of the model for MC scores.

The result that DP and DCP were significant predictors of overall understanding of statistics was not surprising. The literature is saturated with descriptions of the central role that variability plays in statistical thinking (e.g. Wild and Pfannkuch, 1999; Franklin et al., 2007). A possible limitation to this interpretation is that prior studies on the LOCUS assessment revealed high correlation between MC and CR items. However, the current study did not use the same scoring process as the LOCUS project for the CR items. Instances where students did not fully or correctly answer an item as intended by LOCUS but still received a point for showing evidence of understanding of variability were not uncommon. Future studies that utilize both LOCUS CR items and another outside measurement tool for overall understanding of statistics would help to empirically solidify the role of variability.

Summary

This chapter has presented the results from the analysis of the collected data on students understanding of variability. With this information, one should have the ability to interpret the results to sufficiently respond to the research questions guiding this investigation. In various instances throughout the results, statistical significance was presented. Due to the large sample size, small deviations from the null value can display significance when the variation in the data is relatively low. However, practical significance of the results is left open to interpretation.

While the results of this study contributed to existing discussions about students' understanding of variability, it also presented one of the first large-scale examinations of understanding at the general perspective level. Students showed significantly more understanding of variability from the design than the data-centric perspective. However, this difference was only six percentage points and may not necessarily be practically significant. On average, students showed moderate evidence of understanding from each of the two perspectives.

The individual cells of the framework yielded a finer resolution for where students were and were not showing evidence of understanding of variability. While DP1 and DCP1 resulted in nearly all students showing evidence of understanding, DP3 and DCP3 had more variability in their outcomes. The results for DP1 and DCP1 suggest that beginning conceptions of variational disposition may be well known among high school students that have taken a statistics course. These students may be more familiar with concepts of statistical and survey questions and have seen enough examples of data to understand that data vary.

Item parts that addressed the variability and relationships among data and variables element of the framework seemed to yield less evidence of understanding of variability, with only 32% of responses providing evidence. The concepts in this element of variability were more challenging and required a deeper understanding of how to control sources of random variability. While some students seemed to recognize familiar situations that required random sampling, more complex ideas such as the purposes of random assignment were less frequently elaborated on. Students also showed less evidence of understanding when presented with a sampling distribution. Their focus was rarely on variation and the meaning of random chance.

In the second part of the study, the results suggested that understanding of variability according to Peters' (2011) framework was a significant predictor of overall understanding of variability for these students, controlling for all the demographic features that data was collected for. However, these interpretations must only be taken as possible evidence of this relationship because the two measurement tools were from the same assessment despite a unique scoring approach. Additionally, the regression model did not explain around 60% of the variation in the overall understanding of statistics, which means there may be other predictors that were not accounted for in this study or the LOCUS MC scores for these students were highly variable.

Chapter 5 of this dissertation will utilize the results presented here to draw more general conclusions about students' understanding of variability, and will further address the practical significance of the findings. Implications of these results on teaching practice, curriculum development, and assessment will also be discussed.

CHAPTER 5
DISCUSSION

The primary purpose of this study was to develop a large-scale snapshot of U.S. high school students' understanding of variability. To do that, it was necessary to define what it means to understand variability and look at prior studies on K-12 students' understanding of variability. Prior research suggested that students did not tend to have a robust understanding of variability in a variety of different contexts (e.g., Ben-Zvi, 2004; Reading & Shaughnessy, 2000; Reading, 2004; Shaughnessy & Ciancetta, 2002; Torok & Watson, 2000). Multiple context specific frameworks have been utilized and developed in previous students. The Framework for Robust Understanding of Statistical Variation (Peters, 2011) was selected for use in this study because it described variation on an abstract level, void of context, and emphasized the different perspectives from which variability exists in statistical investigations. Once the background was set, this research was able to move forward. This chapter discusses the conclusions and implications that resulted from this research.

Form 2 of the Intermediate/Advanced version of the NSF-funded (DRL-1118168) LOCUS assessments (Jacobbe et al., 2014) were sent to secondaryschools in high-performing districts in seven states across the United States. Students that had completed a course involving statistics took the LOCUS assessments during the school day. Their responses to the CR items were examined for evidence of understanding of variability from the design and data-centric perspectives outlined in the Framework for Robust Understanding of Variation. Prior to analyzing their data, each part of the CR items was coded according to the cell of the framework it addressed. Additionally, a scoring procedure was developed as part of this study that described what a response might contain to provide evidence of understanding from a cell of the framework. Item coding was done by both the author of this dissertation and by a faculty

member in statistics education. Item scoring was conducted by the author of this dissertation and a second scoring was completed on a subset of the data by a graduate student in statistics education. Through the LOCUS assessments, coding, and scoring procedures, data were collected to address the research questions posed in the first chapter of this dissertation.

Discussion

A description of how a sample of over 600 U.S. secondary students from high-performing schools understood variability was determined in response to research question 1. On average, the students in the sample had a moderate understanding of variability from both the design and data-centric perspectives. A significantly higher proportion of responses to LOCUS CR item parts displayed evidence of understanding of variability from the design perspective, however, with a difference of only about six percentage points, the practical significance of this difference is low. A closer look at the content within each of the two perspectives shed light on the areas of variability that students tended to show the most, and least, evidence of understanding in.

Within the design perspective, students showed the most evidence of understanding of variability from the variational disposition element. In the context of this study, students exceled at creating a survey question that anticipated variability in the responses. Creating statistical and survey questions are an important part of the formulating questions process component of the GAISE framework (Franklin et al., 2007). The results of this study showed that, given a context, students had the ability to create appropriate survey questions that anticipated variability in the responses. However, this study did not provide conclusions for students' ability to pose statistical questions, an area that students of various ages tend to struggle with (Pfannkuch & Horring, 2005; Allmond & Makar, 2010).

In the element of variability and relationships among data and variables from the design perspective, students also showed varying levels of understanding. The DP3 cell of the framework consisted of study design elements that are used to help control variability such as random sampling, stratified sampling, and random assignment of treatments in experimental studies. Students showed moderately strong understanding of random and stratified sampling. delMas et al. (2007) analyzed tertiary students' understanding of random sampling as part of the NSF-funded Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project. Their results showed that after a first course in statistics, about 65% of students could select a plausible random sample from a given population. Research on students' understanding of random sampling tends to emphasize skills involved in conducting the random sample. The LOCUS items used in the current study emphasized the necessity of random sampling as part of the design of a study, and required students to consider the purpose and effects of the sampling process.

Another result that was seen among the item parts representing the DP3 cell of the framework was that context appeared to play a role in how students chose to sample. Prior research yields conflicting evidence on the influence of context on student thinking. In some cases, different contexts did not seem to affect how students approached the problem (e.g. Torok & Watson, 2000). In other cases, effects of context on students' understanding of variability were observed in both positive and negative ways. Ben-Zvi (2004) found that students' familiarity with the context allowed them to fully engage in the problem and work to find solutions. Reading (2004), on the other hand, found that context was distracting students because they failed to connect the problem to statistical methods they had recently learned in class. The results of this dissertation provided statistically significant evidence that context played a role in

how students chose to take a sample from a population. A higher proportion of responses described a method for random sampling on the student council problem than on the department store problem, which included an identical item part. While the responses on the department store problem did not show preference for an alternative type of sampling procedure, many responses did not describe a sampling method.

The parts of the variability and relationships among data and variables element in the design perspective that students failed to show evidence of understanding were in experimental design. The purpose of random assignment was only described in 4% of responses and the type of conclusion that could be drawn without an experimental design was only correctly identified in 14% of responses. In the item part that addressed random assignment, the wording of the question may not have persuaded students to explain, in depth, that random assignment is used to help eliminate the effects of variability from variables not controlled in a study. Therefore, the 4% may be lower than the number of responses that would show evidence of understanding if the question prompted more explanation. Random assignment of treatments is a concept that students are not expected to fully understand until they are operating at Level C of the GAISE framework (Franklin et al., 2007). Therefore, we would expect high school students that have taken a course involving statistics to be familiar with the concept of random assignment, but would not yet have mastered the role that variability plays in the process.

The lack of responses displaying evidence that students understood when cause and effect conclusions were appropriate could also be attributed to low levels of understanding of statistics in the sample. The interpretation of the results for this item were more direct than the item that addressed random assignment because responses that were scored as 0 also made inappropriate conclusions from the given data. The 14% of responses that showed evidence of understanding

of variability for cause and effect conclusions were ones that correctly identified that a cause and effect conclusion was inappropriate for the scenario. Distinguishing between conclusions from association studies and experiments is included in the description of the Interpreting Data component of Level C in the GAISE framework. Therefore, students that have only completed an introductory course on statistics would not necessarily be expected to have a comprehensive understanding of appropriate conclusions.

LOCUS CR item parts that addressed variability from the design perspective yielded polarizing results. The data suggests that most high school students that have taken a course involving statistics recognize situations where random sampling is necessary, however, the methods used in this study did not allow for the depth of that understanding to be explored. Additionally, most students could produce a survey question that anticipated variability in its responses, a viable skill when conducting a statistical study. On the other hand, item parts that addressed more complex concepts of study design yielded responses with less evidence of understanding of variability. Students failed to provide in depth explanations of the purpose of randomly assigning treatments in an experimental study, although many seemed to recognize random assignment as an important idea. In the scenario where a formal experimental design was not used, most students' responses still concluded that the association in the data could be used as evidence of a cause and effect relationship between the variables. Despite the high frequency of responses that displayed evidence of understanding from the variational disposition element of the design perspective, there is a clear lack of deep understanding among these students of the role variability plays in the design of studies.

Within the data-centric perspective, students also showed the most understanding of variability from the variational disposition element. In the context of this study, students were

required to create a hypothetical data set with responses to their survey question. Their resulting data display needed to show variation in the responses to display their anticipation of variability based on the context of the data. Nearly all the responses in this study displayed evidence of understanding of the variability that would occur in the collected data. Prior research on students' ability to anticipate variability in a sample taken from a population revealed that students had difficulty with this task when given a population distribution. Shaughnessy et al. (1999) and Reading and Shaughnessy (2000) both found that their small sample of students of various grade levels had difficulty estimating the amount of spread that would occur when taking a sample from the given population and were not able to describe spread in sophisticated terms. These studies did reveal that students tended to know that all the draws from the population would not be the same, a finding that is consistent with results from this dissertation.

In the element of variability and relationships among data and variables from the data-centric perspective, student responses provided weak evidence of understanding. This cell of the framework addressed the exploration of controlled and random variability to infer relationships among data and variables. Consistent with Reading and Shaughnessy (2000), students' responses failed to adequately explain variability resulting in only 31% of responses displaying evidence of understanding. In the next item part, evidence of understanding of variability from this cell was identified by an adequate description of the meaning of a p-value. Only 7% of responses showed evidence of understanding that the outcome observed in the scenario was reasonably due to random variation. The concepts addressed by these CR item parts fall under Level C of the GAISE framework. High school students who have taken a course involving statistics are unlikely to have mastered the ideas and vocabulary necessary to fully explain the scenario

presented in the item. However, a strong understanding of variability combined with the scaffolding provided in the item should have enabled students to provide informed responses.

The item parts that addressed the data-centric perspective also yielded polarizing results. Students exceled at anticipating variability that would occur during data collection from a survey, but struggled to show evidence of understanding when exploring variability within a dataset to draw conclusions about the data. The scenario for the item parts that students struggled with may have been unfamiliar to many students, since it involved a simulated sampling distribution. The context may have distracted students from using their comprehension and problem-solving experiences to adequately respond to the item. Unfortunately, there were no other item parts that covered the necessary cells from the data-centric perspective to make comparisons regarding the influence of context. Despite this limitation, there was a lack of evidence that suggested these high school students had a firm grasp on the concept of variability as it applies to a distribution of data.

The exploration that was guided by research question 1 led to multiple conclusions. Students showed strong evidence of understanding from the element of variational disposition across both design and data-centric perspectives. When more direct understanding of variation was required, like when drawing conclusions from data, there was very little evidence of understanding presented. The items that aligned well with the variational disposition element tended to require less explanation to be scored as showing evidence of understanding. For example, stating that a random sample is an appropriate sampling method was enough to show understanding for DP1. The concept of a random sample is frequently repeated throughout a first course in statistics, but students may not have a complete understanding of the reasons for

collecting random samples. Therefore, it is unknown how deeply students in this study understood those concepts.

Understanding of variability in this study was considered from only four of the twelve total cells in the Framework for Robust Understanding of Statistical Variation (Peters, 2011). Of the pieces of the framework from the design and data-centric perspective, it is worth discussing the importance of the four cells represented by LOCUS CR items. The content in the LOCUS assessments was developed and validated by multiple experts in the field of statistics education and the resulting CR items failed to address half of the framework from these two perspectives. Because of the rigorous development process that led to the LOCUS assessments, it is reasonable to assume that the four cells represented by the CR items tend to appear more commonly in the statistical problem-solving process. Alternatively, an analysis and coding of the MC items may reveal that more of the framework is represented throughout the LOCUS assessments beyond the limited number of CR items. Despite the lack of complete framework coverage by CR items, the results of this study still make a case that this sample of students does not have a strong understanding of variability. The cells that were most emphasized require knowledge of variability and how it relates to the context of and relationships between data and variables.

Further, it cannot be stressed enough that the students that participated in this study were from high-performing districts according to standardized tests. These students displayed weak evidence of understanding of variability, especially from the element of variability and relationships among data and variables. These results raise concerning questions not only regarding the quality of statistics instruction that the students in this study have received, but also the quality, or existence, of the statistics instruction that students in schools from average or underperforming districts. Chances are, if the students with some of the best opportunities for

success are struggling to understand a fundamental concept in statistics, the rest of the secondary students in the United States are as well.

In response to research question 2, a multiple linear regression model was fit to the data and revealed that understanding from the design and data-centric perspectives were each statistically significant predictors of overall understanding of statistics. This finding is consistent with statistics education literature that places variability at the center of the subject of statistics. The GAISE framework (Franklin et al., 2007) explicitly recognizes the role of variability within each component of the statistical problem-solving process. In the Wild & Pfannkuch (1999) model for statistical thinking, consideration of variability is noted as a fundamental type of thinking. Even for those who only seek to develop statistical literacy in their students, an understanding of the role of variability in statistical investigations is a crucial component (Moore, 1990; Cobb & Moore, 1997; Shaughnessy, 1997; Moore, 1998; Garfield & Gal, 1999; Gal, 2004).

This study provides one of the first attempts to empirically examine the relationship between an understanding of variability and an overall understanding of statistics. While it is clear that the results of this study suggest a significant relationship between the two, further research is required to confirm the finding. Psychometric analysis of the LOCUS assessment has shown that all of the items (MC and CR) are highly correlated. However, since a different procedure was used to score the CR items in this study, the strong correlation between all of the items may not hold. Future studies that utilize another instrument for either understanding of variability or overall understanding of statistics would allow for more robust comparisons to be made.

Implications

Students' understanding of variability has been studied in various ways, using different frameworks, with students of all ages. This study was one of the first to gather a large-scale snapshot of students' understanding of variability. Watson et al. (2003) used a sample of 746 students in grades 3, 5, 7 and 9 in Tasmanian schools to devise a questionnaire to be used to measure students' understanding of statistical variation. The current study was the first of this scale to be conducted with U.S. high school students that analyzed patterns and trends in understanding across students.

The use of the LOCUS assessment's CR items to look at an individual concept of statistics provided a new way of utilizing the exam in a research setting. The CR items analyzed as part of this study had a focus on features of study design and the role variability plays in the design of studies and experiments. Previous research tended to focus on elements of variability such as its meaning, anticipating it, and exploring it in data and sampling situations, but rarely considered how students understood the role of variability in the design of studies. In light of the unique features of this study, the implications for curriculum changes, teaching practice, and future research will be discussed in the following sections.

Implications for Curricula

The results of this study suggest that a review of the prominence of the concept of variability in current statistics curricula may be necessary. Students showed the ability to present commonly used phrases like "random sampling" and displayed beginning level data investigation skills like creating a survey question and anticipating variability in a hypothetical set of data. However, evidence of understanding of the role variability plays in the design of studies and experiments, and data exploration and inference was nearly non-existent. The literature presented in this study strongly emphasized the importance of variability as a fundamental concept to

statistics. Thus, statistics curricula should reflect that importance to change the way students approach statistical investigations.

Randomization is an integral part of the study design process and is used in both sampling methods and experimental design, among other places. The random assignment of treatments in an experimental study is necessary to reduce variability in the observed data that results from variables not controlled in the study. Student responses in this study only showed evidence of understanding the purpose of random assignment of treatments 4% of the time, even though many responses acknowledged the idea of random assignment. This may suggest that most students do not fully understand the reasoning behind random assignment of treatments, which acts as a way of limiting the effects of random variation.

A closer look at some commonly used curricular guidelines revealed a lack of emphasis on the concept of variability in experimental design. For example, the Common Core State Standards for Mathematics (NGACBP & CCSSO, 2010), which were adopted by or influenced standards in at least 42 states, only requires students to "use data from a randomized experiment to compare two treatments." Nowhere in the standards are students expected to connect the idea of randomized experiment back to variation. The most recent Principles and Standards for School Mathematics document (NCTM, 2000) expects students to "know the characteristics of well-designed studies, including the role of randomization in each" (p. 324). This expectation emphasizes the role of randomization, but does not tie in the concept of variation. The clear lack of high school students' ability to show evidence of understanding of a commonly covered topic like random assignment of treatments suggests that another look at the focus of statistics curricula may be necessary.

Another important topic in introductory statistics courses is the types of conclusions that can be drawn from studies. Responses in this study showed weak evidence of understanding of the types of conclusions that can be drawn from a study with no experimental design. Only 14% of responses stated that cause and effect conclusions could not be made from a correlation between two variables. This result again suggests that curricular materials may not focus enough on the underlying role of variability. For example, in the Principles and Standards for School Mathematics (NCTM, 2000) data and probability strand for students in grades 9-12, one standard states that all students should "understand the difference among various kinds of studies and which types of inferences can legitimately be drawn from each" (p. 324). The instructor can either explain various types of studies and have students memorize the types of inferences that can be drawn, or the instructor can have students explore the role of variability in each type of study, where and how it is or is not controlled for, and infer the types of conclusions that can be drawn. While standards cannot necessarily be criticized for the resulting teaching methods, they can offer more explicit emphasis on variability because of its importance to the subject of statistics.

This study on high school students understanding of variability presented evidence that students were able to anticipate variability, but struggled with acknowledging, accounting of, and allowing for variability. Statistics curricular materials need to be re-evaluated to consider more explicit expectations for understanding variability and the roles that it plays throughout the statistical problem solving process. The GAISE framework (Franklin et al., 2007) considers the three areas of variability that students struggled with in this study to be central to the collecting, analyzing, and interpreting data components of the statistical problem-solving process, respectively. When developing or making changes to existing statistics curricula, writers should

keep variability in mind to stress the importance of its role in understanding of statistics. The continued used of the GAISE framework as an influential document when designing or improving courses could result in the proper emphasis on understanding of variability.

This dissertation suggests that many high school students do not have a deep understanding of variability after taking courses involving statistics. One possible implication is that statistics courses are failing to provide students with the proper instruction to develop statistical literacy. To fully emphasize the role of variability within the statistical problem-solving process, it may be worth exploring new types of courses that are centered around variation. These courses could provide students with a strong fundamental understanding of statistics at the introductory level in high school or as part of various college programs, such as biostatistics, psychology, or economics. Meletiou and Lee (2002b) developed an introductory course at the college level to test their conjecture that an emphasis on students' intuitions about variability would improve their comprehension of statistical concepts. Although their study focused on a small group of students, the findings suggested improved reasoning about statistical processes. Further investigations on the effects of statistics courses centered around variability are needed to push this agenda, however, there is a clear foundation of research showing a lack of understanding of variability among students.

Implications for Teaching

Despite some of the preceding criticisms of statistics curriculum documents, they do explicitly focus on ideas that are directly linked to the concept of variability. The implementation of curricular materials determines the amount of attention paid to variability in the classroom. The results of this research have implications for teaching practice that could lead to students having a more robust understanding of variability throughout the statistical problem-solving process.

The results of student understanding of variability from the design and data-centric perspectives suggest that changes to the way statistics is taught may be necessary. Less than half of all responses showed evidence of understanding of variability, which could imply a lack of focus on the concept during statistics instruction. Explicitly noting variation in the classroom when appropriate is a simple way to ensure that classroom discussion involves the concept. This suggestion is in line with prior research on student understanding of variability that has also called for changes in the way statistics is taught (e.g. Torok and Waton, 2000; Moore, 1990).

The results from the school day problem suggest that students are not getting enough experience with certain types of data. For example, only 19% of students showed evidence of understanding when faced with exploring and explaining the variation in the simulated data set in the school day problem. Giving students more time and experiences exploring data distributions will allow them to develop their ability to describe what they see. Many responses failed to adequately explain the simulated results by addressing the chance variation seen in the distribution. Discussions of data should not only include measures of central tendency, but also incorporate spread, the shape of the data, and different types of variability (Reading & Shaughnessy, 2004).

The hearing loss problem also revealed high school students' lack of experience with data-based investigations. Only 14% of students noticed that important sources of variability were not controlled through the design of the study and causal conclusions could not be made from the given data. At the same time, many students correctly identified and described the correlation that appeared in the data. Allowing students to investigate more scenarios involving data may give them the experience needed to fully appreciate the entire problem-solving process, and make connections between study design and the types of conclusions that are appropriate.

Advances in technology provide students with the ability to carry out computations that were once burdensome, and allow students to explore real-world scenarios that illustrate variability (Reading & Shaughnessy, 2004; Torok & Watson, 2000). The use of software in the classroom would allow students taking applied statistics courses to more easily explore data in the setting of the class. Thus, there is ample opportunity for more realistic data investigations in the statistics classroom.

Statistics courses have made great progress in their focus for developing conceptual understanding of statistics with the help of the GAISE framework, NCTM Principles and Standards, and the CCSSM. The lack of evidence of understanding in the element of variability and relationships among data and variables suggests that certain topics may require different approaches. Only 32% of responses showed evidence of understanding on items that addressed experimental design, appropriate inferences, and data exploration. By keeping variability as the foundational concept driving all statistical methods and techniques, students will have more exposure to the multitude of roles that variability plays. Research on teaching practices supporting the concept of variability in the classroom is necessary to determine optimal approaches.

Through the implementation of suggestions like these, students may have a more complete understanding of variability from various perspectives. Thinking of study design and data exploration starting with a consideration of variability may solidify many of the concepts commonly taught in statistics courses. While extra emphasis on variability may intrude on time usually spent on other statistical ideas, the central role of variability to statistics supported through the findings of this study and described in the literature provide justification for action.

Implications for Future Research

The Framework for Robust Understanding of Variation (Peters, 2011) is a powerful tool for examining understanding of variability. The current study only utilized the most basic features of the framework to describe the responses seen on LOCUS CR items, and can be used as a guide for further studies. Peters (2011) suggests that the framework be used as a tool for research on instruction, classroom discourse, curriculum, and assessing the effectiveness of textbooks and other materials. The framework allows for connections to be made both within and across the multiple perspectives from which variability can be approached.

Analyzing high school students' understanding of variability with LOCUS CR items resulted in some challenges. Since the items on LOCUS were not written using the Framework for Robust Understanding of Variation (Peters, 2011), they did not always prompt students to specifically address variability. Students' responses could only be interpreted as containing evidence of understanding of variability or not containing evidence. Conclusions were limited in this regard, since it could not be determined if students did not have an understanding of variability. It may be useful to develop further assessment tools that explicitly prompt students to explore variability in the items' scenarios.

The quantitative methods utilized in this study allowed for analysis of patterns and comparisons across a large number of high school students. While conducting the scoring procedure, it became clear that a qualitative study that focused on the content of individual responses on LOCUS CR items would provide researchers with quality information about how students understand variability and other concepts in statistics. For example, the comparison of the two identical item parts with different contexts in this study could be explored further by examining how students responded to each of the two items. Reading and Shaughnessy (2000) and Shaughnessy et al. (1999) recognized that students tended to avoid using words like "vary"

"deviate" and "variation" when describing spread. Analyzing how students responded to LOCUS CR items could help contribute to research on the type of language students use when discussing variability.

Despite the limitations of this study, the use of LOCUS CR items to analyze students' understanding of variability would be useful to formatively assess students taking introductory statistics courses. In fields where students are often required to take statistics courses, such as biostatistics or psychology, an initial understanding of their experiences with variability could be useful to the instructor. Not only would it provide details on their prior knowledge of statistics and variability, it would allow the instructor to adjust course materials to emphasize areas of weakness. Alternatively, using the methodology from this study as an end of course assessment to determine how students understand variability after instruction would provide useful feedback to the instructor.

Conclusion

Overall, this dissertation helped develop a starting point for how U.S. secondary students in high-performing districts understand the concept of variability and contributed empirical evidence that understanding of variability plays a significant role in overall understanding of statistics. Throughout this study, many questions were raised that will require future research to help the field fully understand how students understand variability. The results of this and future studies will assist in the development of plans of action to ensure students from all areas are walking away from statistics courses with a coherent way of thinking about data in the face of variation. Despite the limitations of this research, the findings suggest that even some of the best students' understanding of variability is not yet robust across the board, and that is an important area of research worth pursuing.

APPENDIX A
INFORMED CONSENT LETTER

Informed Consent

Please read this consent document carefully before you decide to participate in this study.

Purpose of the Research Study:
The purpose of this study is to explore what students understand about statistics. This project is developing an assessment that will eventually be used to help improve the way statistics is taught in schools.

What students will be asked to do in the study:
Students will be asked to complete a written test related to statistics content and a survey related to how much exposure they have had to statistics.

Time required:
Each written test will take approximately 90 minutes (broken up into two 45 minute sessions) for students to complete.

Risks and Benefits:
There are no risks associated with participating in this project.
Potential benefits include exposure to statistics concepts more related to everyday life.

Compensation:
Each student will receive a $5 gift card for returning a consent form signed by the student and her/his parent/guardian. Students who choose to participate in the study as well as those who choose not to participate in the study are eligible for the gift card.

Confidentiality:
Student identity will be kept confidential to the extent provided by law. Student information will be assigned a code number. Student names will not be used in any report.

Voluntary Participation:
Your participation in this study is completely voluntary. There is no penalty for not participating.

Right to withdraw from the study:
You have the right to withdraw from the study at anytime without consequence.

Whom to contact if you have questions about the study:
Tim Jacobbe, Assistant Professor, School of Teaching and Learning, University of Florida
PO Box 117048, 2403 Norman Hall, Gainesville, FL 32611; ph: 352-273-4232; e-mail: jacobbe@coe.ufl.edu

Whom to contact about your rights as a research participant in the study:
UF IRB Office, PO Box 112250, University of Florida, Gainesville, FL 32611-2250; ph: 352-392-0433

Signatures: (*Please place an X on the appropriate lines.)*

____ I have read the procedure described above. I voluntarily **give my consent** for my child, _____, to participate in the study. I have received a copy of this description.

____ I have read the procedure described above. I **do not give my consent** for my child, _____, to participate in the study. I have received a copy of this description.

Parent/Guardian: _____ Date: _____

APPENDIX B
SCORING PROCEDURE FOR CR ITEMS

Item 1: Department Store -

Part (a) - Score a 1 for DP3 if response addresses controlling for variability in the study design using random selection. Score a 0.5 if the response addresses an appropriate stratification technique, but does not address randomness. Otherwise, score a 0.

Notes: random selection is used to select the sample, and/or a method is described for how this could be done by numbering the credit card holders from 1 to 5300. Student only needs to specify random selection. If the student specifies random selection, but then describes a process in detail that would NOT result in a random selection, no points given for random selection.

Part (b) - Score a 1 for DP3 if response explains why random assignment to treatments is important in the design of a statistical experiment. Otherwise, score a 0.

Notes: indicates that flipping a coin to determine which advertisement is read results in random assignment to treatments and therefore the study is able to conclude cause and effect. Response must conclude something about the purpose of random assignment. If the response alludes to the idea of confounding variables or cause and effect conclusions, a 1 can be given.

Item 2: Student Council

Part (a) - Score 1 for DP1 if the response contains a survey question that anticipates variability in the responses. Otherwise, score 0.

Notes: Response is a question that asks a student which of the listed activities they prefer. The survey question must be relevant to the context.

Part (b) - Score 1 for DP3 if the response describes a reasonable way to select a random sample of students. Score a 0.5 if the response addresses an appropriate stratification technique, but does not address randomness. Otherwise, score 0.

e.g., recognizes the need for random sampling and clearly describes a process that would result in a random sample from an appropriate population, or the response may stratify the population (divide into subgroups) based on sex or grade level and then select a random sample from each of these groups

Part (c) - Score 1 for DCP1 if the response shows anticipation of reasonable variability in the data by considering the context.

e.g., response creates a table or graph where hypothetical responses are in more than one category. The graph or table does not necessarily have to be completely accurate as long as it's clear that the response anticipates variability in the data.

Part (d) - Rare scenarios may consider variability but problem does not inherently address the concept. Therefore, do not score.

Notes: There were no responses that showed evidence of understanding of variability on this part.

Item 3: Boss Preference

Do not score. This item does not require students to understand variation from the design or data-centric perspectives.

Item 4: Hearing Loss

Part (a) - Does not require understanding of variation. Do not score.

Part (b) - Requires understanding of variation from the modeling perspective. Do not score.

Part (c) - Score 1 for DP3 if the response that indicates that it is not reasonable to conclude that listening to music at high volume is the cause of hearing loss and provides an explanation that is linked to the study design (lack of experimental design). Otherwise, score 0.

Notes: Response states that this was not an experimental study/this was an observational study, or mentions the possibility of confounding variables (i.e., variation was not well controlled). Stating that it was not reasonable to conclude cause and effect due to a small sample size did not receive a 1. If student alludes to interference from outside variables that were not controlled for, score a 1. A very small number of responses showed evidence of understanding from other cells of the framework. These responses were recorded, but since there were no other similar responses, they were still scored a 0 for DP3 and were not within the focus of this study.

Item 5: Extended School Day

Part (a) - Does not consider variation from the design or data-centric perspectives. Do not score.

Part (b) - Score 1 for DCP3 if the response acknowledges that the sample means in the dotplot vary.

Notes: Response indicates counting the number of responses as or more extreme than 24% or explicitly discusses variability in the shown sampling distribution. Any response that acknowledged appropriate variability, for example by counting points or shading, around the 24% mark on the plot received credit.

Part (c) - Score 1 for DCP3 if the response acknowledges sampling variability in their justification of their answer. Otherwise, score 0.

e.g., Response indicates that the difference between the observed sample percentage of 24% and the hypothesized percentage of 30% is not statistically significant or that it could be explained by sampling variability/random chance alone.

LIST OF REFERENCES

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, *21*(1), 243-247.

Allmond, S., & Makar, K. (2010, July). Developing primary students' ability to pose questions in statistical investigations. In *Proceedings of the 8th international conference on teaching statistics. Voorburg, The Netherlands: International Statistical Institute*.

American Community Survey. (2015). Detailed Languages Spoken at Home and Ability to Speak English for the Population 5 Years and Over: 2009-2013. Retrieved from https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html.

American Statistical Association. (1991). Guidelines for the teaching of statistics K-12 mathematics curriculum. *Alexandria, VA: Author*.

Batanero, C., Cobo, B., & Díaz, C. (2003). Assessing secondary school students' understanding of averages. In *Proceedings of CERME* (Vol. 3).

Ben-Zvi, D. (2002). Seventh grade students' sense making of data and data representations. In *Proceedings of the Sixth International Conference on Teaching of Statistics*.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42–63.

Ben-Zvi, D., & Garfield, J. (2004). Research on reasoning about variability: A forward. *Statistics Education Research Journal*, *3*(2), 4–6.

Bidell, T. R., & Fischer, K. W. (1992). Beyond the stage debate: Action, structure, and variability in Piagetian theory and research. *Intellectual development*, 100-140.

Biggs, J. (1989). Towards a Model of School-Based Curriculum Development and Assessment Using the SOLO Taxonomy. *Australian journal of education*, *33*(2), 151-63.

Biggs, J. B., & Collis, K. F. (1982). *Evaluation the quality of learning: the SOLO taxonomy (structure of the observed learning outcome)*. Academic Press.

Biggs, J., & Collis, K., (1991). Multimodal learning and the quality of intelligent behavior. In H. Rowe (Ed.), *Intelligence, Reconceptualization and Measurement* (pp. 57-76). New Jersey: Laurence Erlbaum Assoc.

Case, C., & Jacobbe, T. (2014). Lessons from the LOCUS assessments: Center, spread, and informal inference. *The Statistics Teacher Network*, 84, 14-17.

Case, R. (1985). *Intellectual development: Birth to adulthood*. Academic Pr.

Case, R. (1992). *The Mind's Staircase: Exploring the conceptual underpinnings of children's thought and knowledge*. Hillsdale, NJ: Erlbaum.

Case, R., & Okamoto, Y. (1996). The role of central conceptual structures in the development of Gifted Learners' Epistemological Beliefs 65 children's thought. *Monographs of the Society for Research in Child Development*, 61(1-2, Serial No. 246).

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

Chance, B., delMas, R. & Garfield, J. (2004). Reasoning about sampling distributions. In The Challenge of Developing Statistical Literacy, Reasoning, and Thinking, Eds. D. Ben-Zvi and J. Garfield, pp. 295–323. Dordrecht: Kluwer Academic.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801-823.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Collis, K. F., & Biggs, J. B. (1991). Developmental determinants of qualitative aspects of school learning. *Learning and teaching cognitive skills*, 185-207.

Cruz, J., & Garrett, A. J. (2006). Students' actions in open and multiple-choice questions regarding understanding of averages. *International Group for the Psychology of Mathematics Education*, 161.

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79–95). Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/1-4020-2278-6_4.pdf

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55-82.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method*. John Wiley & Sons.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological review*, *87*(6), 477.

Foti, S., & Jacobbe, T. (2015). Lessons from the LOCUS assessments: Boxplots. *The Statistics Teacher Network*, 85, 2-5.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: a pre-K--12 curriculum framework.* Alexandria, VA: American Statistical Association.

Gal, I. (ed.), (2000), *Adult Numeracy Development: Theory, Research, Practice*. Cresskill, NJ: Hampton Press.

Gal, I. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review / Revue Internationale de Statistique, 70*(1), 1. https://doi.org/10.2307/1403713

Gal, I. (2004). Statistical literacy. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 47–78). Springer. Retrieved from http://link.springer.com/chapter/10.1007/1-4020-2278-6_3

García Cruz, J. A., & Garrett, A. J. (2008). Understanding the arithmetic mean: A study with secondary and university students. *Journal of the Korea Society of Mathematical Education Series D: Research in Mathematical Education*, *12*(1), 49-66.

Garfield, J. (1999). Thinking about statistical reasoning, thinking and literacy. *First Annual Roundtable on Statistical Thinking, Reasoning and Literacy (STRL-1).*

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Science & Business Media.

Garfield, J. B., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review*, *67*(1), 1-12.

Jacobbe, T. (2015). NSF funds research in statistics education: LOCUS. Amstat News, 35-36.

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the content validity of the LOCUS assessments through evidence centered design In K. Makar & R. Gould (Eds.) Proceedings of the 9th International Conference on Teaching Statistics.

Jacobbe, T., Whitaker, D., & Foti, S. (2015). The Levels of Conceptual Understanding in Statistics (LOCUS) Project: Results of the Pilot Study. *Numeracy*, *8*(2), 3.

Jones, G. A., Langrall, C.W., Thornton, C. A., & Mogill, A. T. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educational Studies in Mathematics, 32,* 101–125.

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A Framework for Characterizing Children's Statistical Thinking. *Mathematical Thinking and Learning*, *2*(4), 269–307. https://doi.org/10.1207/S15327833MTL0204_3

Konold, C., & Pollatsek, A. (2002). Data Analysis as the Search for Signals in Noisy Processes. *Journal for Research in Mathematics Education*, *33*(4), 259. https://doi.org/10.2307/749741

Kruskal, W. H., & Wallman, K. K. (1982). Federal Statistics and You. *Amstat News*, (85), 3-4.

Langrall, C. W., & Mooney, E. S. (2002, October). The development of a framework characterizing middle school students' statistical thinking. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*.

Lann, A., & Falk, R. (2003). What are the clues for intuitive assessment of variability? In *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-3).*

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198-1202.

Makar, K., & Confrey, J. (2003). Chunks, Clumps, and Spread Out: Secondary Preservice Teachers' Informal Notions of Variation and Distribution. *Reasoning about Variability: A Collection of Current Research Studies*.

Mathews, D., & Clark, J. (1997). Successful students' conceptions of mean, standard deviation, and the Central Limit Theorem. In *Midwest Conference on Teaching Statistics, Oshkosh, WI*.

Meletiou, M., & Lee, C. (2002a). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. In *Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa*. Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/1/10_19_me.pdf

Meletiou, M., & Lee, C. (2002b). Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistics Education Research Journal*, 1(2), 22-37.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. *Handbook of test development*, 61-90.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 20-39.

Mooney, E. S. (2002). A Framework for Characterizing Middle School Students' Statistical Thinking. *Mathematical Thinking and Learning*, *4*(1), 23–63.

Moore, D. (1992). Statistics for all: Why? What and how? In D. Vere-Jones (Ed.), Proceedings of the Third International Conference on Teaching Statistics, Vol. 1 (pp. 423–428). Voorburg, The Netherlands: International Statistical Institute.

Moore, D. S. (1990). Uncertainty. *On the shoulders of giants: New approaches to numeracy*, 95-137.

Moore, D. (1998). Statistics among the liberal arts. *Journal of the American Statistical Association*, *93*(444), 1253-1259.

Moore, D. S., & Cobb, G. W. (2000). Statistics and mathematics: Tension and cooperation. *The American Mathematical Monthly*, *107*(7), 615-630.

National Center for Education Statistics. (2014). State Nonfiscal Survey of Public Elementary and Secondary Education. Retrieved from https://nces.ed.gov/programs/digest/d16/tables/dt16_203.50.asp?current=yes.

National Center for Education Statistics. (2015). *2015 Digest of education statistics.* Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_203.10.asp

National Council of Teachers of Mathematics (Ed.). (2000). *Principles and standards for school mathematics* (Vol. 1). National Council of Teachers of.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. National.

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

Piaget, J. (1954). Language and thought from a genetic perspective. *Acta Psychologica*, *10*, 51-60.

Piaget, J. (1962). Commentary on Vygotsky's criticisms of Language and thought of the child and judgment and reasoning in the child. *Lev Vygotsky, Critical Assessments*, *1*, 241-260.

Pegg, J. & Davey, G. (1998). A synthesis of Two Models: Interpreting Student Understanding in Geometry. In R. Lehrer & C. Chazan, (Eds.), Designing Learning Environments for Developing Understanding of Geometry and Spac. (pp.109–135). New Jersey: Lawrence Erlbaum.

Peters, S. A. (2009). Developing an understanding of variation: AP statistics teachers 'perceptions and recollections of critical moments (Doctoral dissertation, Pennsylvania State University).

Peters, S. (2011). Robust understanding of statistical variation. *Statistics Education Research Journal*, *10*(1), 52–88.

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process. *Statistics Education Research Journal*, *5*(2), 4–9.

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 Statistics Education 2004 Roundtable, (pp. 204-218). Voorburg, The Netherlands: ISI.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17–46). Springer. Retrieved from http://link.springer.com/chapter/10.1007/1-4020-2278-6_2

Piaget, J. (1954). Language and thought from a genetic perspective. *Acta Psychologica, 10*, 51-60.

Piaget, J. (1962). Commentary on Vygotsky's criticisms of Language and thought of the child and judgment and reasoning in the child. *Lev Vygotsky, Critical Assessments, 1*, 241-260.

Prodromou, T., & Pratt, D. (2006). The role of causality in the co-ordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, *5*(2), 69-88.

Professional Development. (n.d.). Levels of Conceptual Understanding of Statistics website. https://locus.statisticseducation.org/professional-development

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media.

Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal*, *3*(2), 84–105.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal*, *5*(2), 46–68.

Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In *PME CONFERENCE* (Vol. 4, pp. 4–89).

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi (Ed.), *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht: Kluwer Acad. Publ.

Reid, J., & Reading, C. (2005). Developing consideration of variation: Case studies from a tertiary introductory service statistics course. *Proceedings of the 55th Session of the International Statistical Institute,* Sydney, Australia. Voorburg, Netherlands: International Statistical Institute.

Reid, J., & Reading, C. (2008). Measuring the development of students' consideration of variation. *Statistics Education Research Journal*, *7*(1), 40–59.

Robinson, W. S. (1957). The statistical measurement of agreement. *American sociological review*, *22*(1), 17-25.

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. *People in mathematics education*, *1*, 6-22.

Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In *Proceedings of the Sixth International Conference on Teaching Statistics* (pp. 295–312).

Shaughnessy, J. M., & Zawojewski, J. S. (1999). Secondary students' performance on data and chance in the 1996 NAEP. *The Mathematics Teacher*, *92*(8), 713.

Shaughnessy, J. M., Watson, J., Moritz, J. B., & Reading, C. (1999). School mathematics students' acknowledgement of statistical variation. Presented at the Research Pre-Session of the 77th annual meeting of the National Council of Teachers of Mathematics, San Francisco.

Steen, L.A. (2001). *Mathematics and democracy: The case for quantitative literacy* (pp. 9709547-0). L. A. Steen (Ed.). NCED.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, *12*(2), 147–169.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1 - 67.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*, (Edited by M. Cole, V. John-Steiner, S. Scribner, & E. Souberman). Cambridge, MA: Harvard University Press.

Wallman, K. K. (1993). Enhancing Statistical Literacy: Enriching Our Society. *Journal of the American Statistical Association*, *88*(421), 1. https://doi.org/10.2307/2290686

Watson, J. M. (1997). Assessing statistical thinking using the media. *The assessment challenge in statistics education*, 107-121.

Watson, J. M. (2004). Quantitative Literacy in the Media: An Arena for Problem-solving. *Australian Mathematics Teacher*, *60*(1), 34–40.

Watson, J. M., & Moritz, J. B. (1999). The Development of Concepts of Average. *Focus on Learning Problems in Mathematics*, *21*(4), 15-39.

Watson, J. M., & Moritz, J. B. (2000). Developing Concepts of Sampling. *Journal for Research in Mathematics Education*, *31*(1), 44. https://doi.org/10.2307/749819

Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation*, *1*(3), 247-275.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology, 34*(1), 1–29. https://doi.org/10.1080/0020739021000018791

Watson, J. M., Callingham, R. A., & Kelly, B. A. (2007). Students' Appreciation of Expectation and Variation as a Foundation for Statistical Understanding. *Mathematical Thinking and Learning*, *9*(2), 83–130. https://doi.org/10.1080/10986060709336812

Whitaker, D., & Jacobbe, T. (2014). Lessons from the LOCUS assessments: Comparing groups. *The Statistics Teacher Network*, 83, 13-15.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review/Revue Internationale de Statistique*, 223–248.

BIOGRAPHICAL SKETCH

Steven J. Foti was born in 1989 in western New York. He and his two sisters Jeanna and Kristy are the three children of Joseph and Carolyn Foti. Steve graduated from Greece Athena High School in 2008, from Clarkson University in 2011 with a B.S. in applied mathematics and statistics and physics, from the University of Florida in 2013 with a M.S in statistics, and finally from the University of Florida in August 2017 with a Ph.D. in curriculum and instruction. His major area of concentration for his Ph.D. was statistics education.

Throughout his academic career, Steve has remained active in a multitude of teaching and research roles. At Clarkson University, he was a teaching assistant for mathematics and statistics courses. Steve spent his summers as an undergraduate researcher, in both statistics and physics, and as a mentor to elementary students participating in a roller coaster camp as part of New York State's Science Technology Entry Program. At the University of Florida, he continued to act as a teaching assistant in multiple introductory statistics courses and spent multiple semesters as an instructor for both secondary students and undergraduates. In fall 2014, Steve was awarded a graduate school fellowship to support his pursuit of a terminal degree.

Steve is a peer-reviewed author in the field of statistics education. He has multiple articles approved for publication and has served as both a co- and lead author. He has also presented his work at a variety of local, regional, national, and international conferences including the School Science and Mathematics Association, National Council of Teachers of Mathematics, and the International Conference on Teaching Statistics.

In July 2017, Steve accepted a position at the University of Florida as a clinical assistant professor in the Department of Biostatistics.