UNDERGRADUATE STUDENTS' INFORMAL NOTIONS OF VARIABILITY

by

OGUZ KOKLU

(Under the Direction of Jennifer J. Kaplan)

ABSTRACT

A robust understanding of variability is key to deeper conceptualization of other major statistical ideas, but many students have only naive notions of variability. Researchers have identified some of these informal notions, but existing research is limited with regard to how students reason when their preexisting informal notions are not applicable. Therefore, I investigated undergraduate students' reasoning about variability when datasets or distributions to be compared (a) have equal ranges, (b) do not include extreme values, and (c) have approximately the same number of different values; and the ways, if any, providing a context supports students' reasoning about variability in the preceding situations.

Following the premises of the knowledge-in-pieces epistemological perspective (diSessa, 1993), I designed statistical tasks and used them as homework questions. I analyzed students' responses to homework questions following Arnold's (2013) distribution framework. In addition, I conducted two or three task-based interviews with students using the similar statistical tasks. Using Powell, Francisco, and Maher's methodology (2003), I analyzed four of these students' video recorded interviews.

The analysis of the homework data showed that the students addressed variability considerably less frequently than they addressed the shape of a given distribution. In addition, the students often provided limited responses in their homework questions. The interview data showed that three of the participants had informal notions of variability and employed them inconsistently across the tasks. Overall, the students' reasoning about variability was often contingent upon the particular and more prevalent characteristics of the questions on which they were asked to work. Lastly, although the use of contextual information by the interviewed students was minimal, student responses to homework questions suggested that availability of context in a statistical question changed students' choices from among the incorrect answer options.

The study presents multiple directions to frame future research. The most pressing areas are exploring how statistical terms such as *variability* are used in introductory statistics courses, creating practical intervention tasks that could be used to underline the normative meaning of *variability*, and suggesting instructional designs to exploit students' preexisting statistical notions in developing a more robust statistical knowledge.

INDEX WORDS: Understanding variability, Teaching and learning statistics, Undergraduate statistics instruction.

UNDERGRADUATE STUDENTS' INFORMAL NOTIONS OF VARIABILITY

by

OGUZ KOKLU

B.S., Bogazici University, Turkey, 2008

M.S., Bogazici University, Turkey, 2012

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

UNDERGRADUATE STUDENTS' INFORMAL NOTIONS OF VARIABILITY

by

OGUZ KOKLU

Major Professor:	Jennifer J. Kaplan
Committee:	Jeremy Kilpatrick
	Denise Spangler

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

DEDICATION

I dedicate this study to

Allah, my creator, the most merciful

(& Muhammad (May Allah bless him), messenger of Allah, my greatest teacher)

For sending me to the world and honoring me with Islam.

Life would be meaningless without belief.

ACKNOWLEDGEMENTS

First, I would like thank to my major professor, Jennifer J. Kaplan, for believing in my potential and guiding me throughout the years I worked on my dissertation study. Her comments always motivated me and improved the quality of my study. I also would like to say Thank You to my committee members, Denise Spangler and Jeremy Kilpatrick, for their very thoughtful suggestions and feedback throughout the dissertation process. I always believed that my committee members were ready to help me in achieving my best.

I can never appreciate enough the sacrifices of my wife that allowed me to complete this work. She deserves the entire honor I am receiving from the PhD degree. During the past four years, she and our older daughter, Zumranur, spent nights and days together waiting for me to complete a homework assignment, project, paper, or research duties. Fortunately, our younger daughter, Zehra, is too young to recognize what it means to have a father pursuing a PhD.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

Problem Statement 3

Statement of Purpose and Research Questions 4

2 LITERATURE REVIEW 5

Discussion of the Key Concepts 5

Students' Understanding of Variability 9

Students' Informal Notions of Variability 14

Misconceptions Literature on Variability 20

Reasoning about Variability: What Is Missing? 23

Theoretical Framework 28

3 METHODS 36

Research Setting 36

Participants 38

Data Collection 40

Data Analysis 44

Pilot Study 51

4 RESULTS 53

Overview of the Results 53

Results Specific to Each Research Question 62

Additional Findings 86

5 DISCUSSION 99

Summary of Findings 99

Limitations 105

Implications for Further Research 107

REFERENCES 112

APPENDICES

A RECRUITMENT SCRIPT 125

B HOMEWORK QUESTIONS 127

C BACKGROUND QUESTIONNAIRE, OPEN-ENDED QUESTIONS, TASKS 146

LIST OF TABLES

Page

Table 1: Interview Participants..40

Table 2: Overview of Homework Questions..42

Table 3: Variability in the Distribution Framework in Arnold (2013)...........................49

Table 4: Summary of Homework Data 50

Table 5: Summary of the Use of Variability Terms by 1,156 Students 56

Table 6: Percentages and Examples for Each Category of Response in Question 3......68

Table 7: Percentages and Examples for Both Versions of Question 2...........................84

LIST OF FIGURES

Page

Figure 1: Homework question 2 63

Figure 2: Student explanations from less clear to clearer consideration of center 64

Figure 3: Homework question 8 87

Figure 4: Ocean’s method of calculating “mean” for categorical data 90

CHAPTER 1

INTRODUCTION

Statistical understanding, which has become more valuable due to its potential to enhance the quality of decision making (Graham, Pfannkuch, & Thomas, 2009), is highlighted using different terms, the more common ones being statistical thinking (American Statistical Association [ASA], 2005; National Governors Association Center for Best Practices [NGA] & Council of Chief State School Officers [CCSSO], 2010; Moore, 1990; Scheaffer, 2000), statistical reasoning (NGA & CCSSO, 2010; Franklin, 2013; Wild & Pfannkuch, 1999), and statistical literacy (ASA, 2005; Franklin et al., 2007; Gal, 2002, 2004). Although the emphases implied by each of these terms may diverge, a common theme is the need to consider variability when working with data (Graham et al., 2009). Variability is ubiquitous, and as Cobb and Moore (1997) suggested, without variability there would be no such discipline as statistics. Variability is a core concern in statistical investigations (ASA, 2005; Cobb & Moore, 1997; Franklin et al., 2007), often addressed by understanding, quantifying, explaining, or, if possible, reducing it (such as by controlling for other variables in experimental designs) in data (Franklin et al., 2007; Graham et al., 2009, p. 681, Moore, 1990, p. 135; Wild & Pfannkuch, 1999).

Suggesting a definition of variability that applies to all different types of variables is impossible. In the case of categorical variables, variability indicates the frequency by which “observations differ from one another” (Kader & Perry, 2007, p. 2) across the categories of the variable of interest. The variability of a univariate quantitative variable, on the other hand, suggests the degree to which observations deviate from the measure of center. That is to say, variability of univariate quantitative data requires consideration of the central value and its distance from each observation in a distribution of the data. It should be noted that this requirement is not applicable to categorical variables because a typical measure of center does not exist for categorical data. In brief, variability for categorical data focuses on *how often* observations differ (from each other) and variability for quantitative data focuses on *how much* observations differ (from the mean). Thus, when a definition adequately describes variability for categorical variables, it lacks important aspects of the concept in quantitative variables (Kader & Perry, 2007).

Variability is also the unifying element of statistics instruction in Grades K–12 and at the college level (ASA, 2005, pp. 8–10; Kader & Mamer, 2008, p. 38). In addition to its essential role in investigations with data (Leavy & Middleton, 2011), a lack of comprehension of and attention to variability, especially in working with quantitative data, impedes the understanding of other statistical key ideas such as randomness, distribution, sampling, central tendency, and inference (Makar, 2016). Overall, understanding variability of quantitative data is a major goal of teaching and learning statistics (Wild & Pfannkuch, 1999), thus is a key to success in statistics.

Reasoning about variability is crucial but often challenging for students at any grade level. Previous studies (e.g., Garfield, delMas, & Chance, 2007; Lann & Falk, 2003) have shown that students tend to rely on their preexisting “basic notions of variability” (Pingel, 1993, p. 71). When students are asked about variability, their notions are often reflected in action as (a) comparing only ranges across datasets, (b) focusing only on individual values (usually the extreme ones), and (c) exploring the extent to which observations in a distribution are repeated. In brief, students’ overreliance on their preexisting informal notions of variability, and difficulty in attending to how data values cluster around the central value, make reasoning about variability for numerical variables an overwhelming task for them.

Problem Statement

Variability is a key idea in statistics and has a fundamental role in all aspects of data analysis (Leavy & Middleton, 2011), making the understanding of the concept one of the most important goals of statistics instruction (Franklin et al., 2007). The authors of such documents as *Connecting Research to Practice in a Culture of Assessment for Introductory College-Level Statistics Report* (Pearl et al., 2012), *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report* (ASA, 2005), *GAISE Pre-K-12 Framework* (Franklin et al., 2007), *Common Core State Standards for Mathematics (CCSSM)* (NGA & CCSSO, 2010), and *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000) agree that, although reasoning about variability is crucial for understanding and practicing statistics, it is also multifaceted and complex for students.

Students also have difficulty in reasoning about variability within a *context* (i.e., the real-life circumstance or the scenario presented in a statistical question or an investigation). Instead, students form (and continue to rely on) some informal notions—which often reflect a naive and fragmented understanding of the concept—even after receiving instruction specifically on variability (delMas & Liu, 2007; Garfield et al., 2007). Such informal notions as relying on range or the individual values in a given distribution (Confrey & Makar, 2002; Garfield et al., 2007; Shaughnessy, 2007) and confusing what *variability* refers to in the case of categorical and numerical variables (Loosen, Lioen, & Lacante, 1985) usually cause partial or incorrect conclusions regarding variability in statistical investigations. In addition, they impede students' development of a proper conceptualization of the concept itself (Lehrer & Schauble, 2002).

Because informal notions often result in limited attention to the core meaning of variability and are obstacles to the robust understanding of variability, it is essential to investigate how students reason when their informal notions are not sufficient to address variability.

Statement of Purpose and Research Questions

In this study, I focused on undergraduate students' reasoning about variability for quantitative variables. I aimed to investigate how undergraduate students reason about variability when their informal notions of variability are neither applicable nor fruitful because of the particular characteristics of a given problem, situation, or dataset—both with and without a context. Each case listed in the first question below was formulated to contest the aforementioned informal notions; students are not able to apply these notions, or if they are, they do not find them to be helpful. Another area of focus for the study is whether and to what extent there is link between students' informal notions and the presence or absence of a context. I sought answers to the following research questions:

1. How do undergraduate students reason about variability when the datasets or distributions to be compared have
    a. equal ranges?
    b. no extreme values?
    c. approximately the same number of different values?
2. In what ways, if any, does providing a context support or detract from students' reasoning about variability in the preceding situations?

CHAPTER 2

LITERATURE REVIEW

In this chapter, I elaborate on the research focusing on students' understanding of variability, primarily in the case of univariate quantitative data. Although the population of interest in the study was students enrolled in undergraduate introductory statistics courses, I also reviewed studies that targeted middle and high school students' (and occasionally preservice and in-service teachers') understanding of variability. I assume that undergraduate students' reasoning about variability might not be considerably different from that of middle and high school students.

Note that a robust comprehension of variability is challenging without attending to other statistical concepts (Watson & Kelly, 2007) because knowing variability involves understanding many "statistical big ideas" (Garfield & Ben-Zvi, 2005, p. 98; Watson & Kelly, 2007) and their multifaceted relationships with variability (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). Given that understanding variability hinges upon comprehending other statistical big ideas, it is challenging to include every aspect of research on variability in a literature review. Consequently, I reviewed the research studies that are more closely related to my study, which could be summarized as the literature pertaining to undergraduate students' reasoning about variability for univariate quantitative variables.

Discussion of the Key Concepts

The normative statistical meaning of *variability*, as understood by professional statisticians and as traditionally described in textbooks (e.g., Bock, Velleman, & De Veaux, 2012), refers to the measures that indicate *how data values typically deviate from a center value*. It should be noted that this description is appropriate for numerical data, especially univariate quantitative data. The meaning of *variability* for categorical variables is different. Below, I discuss variability for categorical variables before I focus on variability for univariate quantitative variables.

If the outcomes of a variable fall into categories, then the variable is called categorical, and the variability of categorical data indicates how frequently "the observations *differ from one another* [emphasis added]" (Kader & Perry, 2007, p. 2). Kader and Perry (2007) and Perry and Kader (2005) coined the term *unalikeability,* for the description. In their introductory statistics textbook, Gould and Ryan (2014) used the term *diversity* to introduce variability for categorical data. Although the term *diversity* seems to be more self-explanatory, unalikeability and diversity are consistent in ways they refer to variability, both suggesting that the variability of a categorical variable indicates how frequently the observations differ from each other in a dataset. Therefore, if the observations are spread more equally across multiple categories, then the variable of interest is more varied than if only a few observations are so spread. For example, a college with approximately equal numbers of Hispanic, Asian, and African-American students is more *diverse,* hence more variable in terms of the categorical variable *ethnicity,* than a college with a majority of students who are Hispanic but with few students of other ethnicities. In brief, students should recognize the difference between "how often the observations differ from one another" (to measure the variability of categorical variables) and "how much the values differ from the mean" (Kader & Perry, 2007, p. 1) (to measure the variability of quantitative variables).

A variable with a single numerical characteristic is called a *univariate quantitative variable*, and two approaches can be used for the variability of this kind of variable (Jones &

Scariano, 2014, p. 93; Kader & Jacobbe, 2013, p. 25). The first approach is based on the distance between two individual data values (i.e., observations) as in the case of range (Jones & Scariano, 2014). Range, the numerical value of the difference between the maximum and the minimum values of a variable, provides the *overall* spread but ignores (the spread of) all other values in a dataset. In other words, in range, only the most extreme possible difference in a given dataset is considered (Kader & Jacobbe, 2013, p. 25; Tabor & Franklin, 2013, p. 130). Similarly, the interquartile range (IQR), which could be regarded as a refined version of range, ignores the extreme values and takes only the middle 50% of the data into consideration. Overall, comprehending the idea behind range and IQR is simple, and these measures can be useful tools to compare the variability of two or more datasets quickly, but in a simplistic way (Garfield & Ben-Zvi, 2005). The second approach to gauge variability, which is more representative of statistical norms, is based on the average distances of the observations in a dataset from a centrally located value, usually the mean of the observations (Jones & Scariano, 2014; Kader & Jacobbe, 2013). It is important to note that, in this approach, variability is the characteristic of a distribution (Ciancetta, 2007), and each value in the given distribution contributes to the property in different weights (Peck et al., 2013, p. 25): The farther a value from the mean, the greater its contribution to the variability.

The widely used formal *numerical measures* of variability are standard deviation (SD) and variance (Bock et al., 2012, p. 60; Pingel, 1993). Undergraduate students should understand variance as the average of the squared deviations from the mean, and SD as the square root of the variance. Similarly, mean absolute deviation (MAD) and, equivalently, sum of the absolute deviations (SAD), measure variability in a similar way that SD and variance do. MAD and SAD are usually taught in middle and high school (see NGA & CCSSO, 2010; Franklin et al., 2007) but typically not in undergraduate statistics courses, and they are rarely used in the statistics profession (see ASA, 2005).

Standard deviation—and also MAD—characterizes the average spread of the data from the measure of the center (Konold & Pollatsek, 2002). In other words, SD suggests the overall measure of how variable the values are, on average, from the mean (Peck et al., 2013, p. 17). Students should comprehend that most of the values need to cluster more closely around the mean in order for a distribution to have a relatively smaller SD, and thus smaller variability (delMas & Liu, 2007). Statistics instruction should be instrumental in students' discovery that SD is a measure of the density of observations about the mean of a distribution (delMas & Liu, 2005). Makar and Confrey (2005) and Peters (2010) suggested that students usually struggle to interpret these numerical measures in the context in which data are provided.

These two approaches to variability, *distance between two points* and *average distance from the mean*, emphasize different characteristics of distributions, making the conceptualization of variability challenging for students. Measures of variability such as range and IQR are appropriate to use if the aim is to achieve a rough estimation of variability. Although using range and IQR is practical if a dataset has only a few values and if there is no common pattern or clustering around a point in the dataset, these measures are less helpful if a distribution is approximately symmetric and bell-shaped (Bock et al., 2012, pp. 43–82). In addition, having equal ranges tells little about the dispersion of the datasets to be compared. On the other hand, although measures of variability such as SD and MAD prioritize information on how observations differ from the mean, use of these measures is less useful in skewed distributions because the mean and SD are distorted in a skewed distribution. On the contrary, students may erroneously think, "The standard deviation adequately quantifies the variability of every set of scores" (Pingel, 1993, p. 70). To conclude, students need to recognize the difference between these distinct approaches (i.e., measures) to variability and build skills that are helpful in selecting and employing the most appropriate approach under particular conditions (Garfield & Gal, 1999, pp. 210–211).

Understanding the concept of variability for univariate quantitative variables is instrumental in learning the concept as applied to other distribution sources (e.g., in the cases of chance and sampling variability) and types of numerical variables (e.g., variability of bivariate data) (Peck et al., 2013, p. 4). For instance, variability of bivariate data refers to a measure indicating how data typically varies from a *line (or a curve)* (Garfield & Ben-Zvi, 2005; Peck et al., 2013, p. 4). The more data points stray from the line or curve, the larger the variability will be, which constitutes the foundational idea to understand statistical *covariation* (Cobb, 1999).

The scope of research on reasoning about variability is extensive (Lehrer & Schauble, 2004). In this section, I described how I conceptualized the concept of variability and explained how I treat the concept throughout the study. In the next section, I focus on students' understanding of variability.

Students' Understanding of Variability

Although statisticians and textbooks agree on a normative statistical meaning of *variability*, students do not necessarily understand this term in the ways it is used by expert statisticians or the statistical community. Students often have fragmented, incomplete, or even contradictory interpretations of the concept (Garfield et al., 2007; Lann & Falk, 2003; Loosen et al., 1985; Pingel, 1993, p. 71). In an extensive analysis of studies on variability, Shaughnessy (2006, pp. 92–93; 2007, pp. 984–985) identified eight conceptions of variability commonly held by students:

- *Unusual values such as extremes or outliers*: In a given distribution or dataset, students focus their attention on particular data values such as unusual values or tails of a distribution to account for variability.
- *Change over time*: Students observe variability as the repeated measurement of a variable over time.
- *Whole range*: Students focus on the spread of all possible values. They no longer see data as individual points that vary but begin to recognize that a whole dataset can vary.
- *Likely range of a sample*: Students generalize the idea of relative frequency to samples. The variability within or across samples can be brought into focus with this conception, which can also help students understand sampling variability.
- *Difference or distance from a fixed point*: Students perceive variability as measuring the distance between a point and a reference point (such as a mean) and then repeating the same calculation for all observations in a distribution.
- *Sum of residuals*: Students understand variability as the measure of the total amount a distribution is spread out from a fixed value, usually the mean.
- *Covariation or association*: Students regard variability as an interaction of variables such as how a change in one variable may co-occur with a change in another variable.
- *Distribution*: Distributions vary. Students compare variability between or among distributions to help understanding data.

It should be noted that the list may not be exhaustive, and Shaughnessy (2006, 2007) did not attempt to list these understandings in order from simple to more complicated. Variability is a multifaceted notion, and each conception suggested by Shaughnessy may be helpful in particular cases (Garfield & Ben-Zvi, 2005). For example, understanding that *distributions vary* is a crucial observation for students to conceptualize variability. Students may benefit from the notion in understanding the more formal measures and normative interpretations of variability. As Ciancetta (2007, p. 41) and Lehrer and Schauble (2004, p. 638) suggested, students' understanding of variation is directly related to their understanding of the overarching concept of distribution. Hence, in the following section I illustrate this relationship by elaborating on the unifying concept of *distribution*.

Shape, Center, and Variability of a Distribution

The conceptual understanding of the *shape*, *center,* and *variability* of a *distribution* are linked (Ciancetta, 2007; delMas & Liu, 2005; Leinhardt & Larreamendy-Joerns, 2007, p. 187; Makar & Confrey, 2003; Reading & Reid, 2006). Distribution refers to "the arrangement of the observations along a scale of measurement" (Hardyck & Petrinovichas as cited in Leavy & Middleton, 2011, p. 235). A *distribution of a variable* is a visual representation that aids to show outcomes of a variable and the relative frequency of these outcomes (Bock et al., 2012; Tabor & Franklin, 2013, p. 120). A distribution of quantitative data reveals important features of the data through the distribution's overall shape, location of its center, and its variability in the given context of the data (delMas & Liu, 2005).

Understanding the statistical concept of distribution is an essential component of statistical reasoning (Arnold & Pfannkuch, 2014; Franklin & Kader, 2006, p. 1; Garfield & Ben-Zvi, 2005; Lehrer et al., 2011). Studies on students' understanding of distribution (e.g., Arnold, 2013; Arnold & Pfannkuch, 2014; Bakker & Gravemeijer, 2004; Ben-Zvi & Arcavi, 2001; Garfield & Ben-Zvi, 2008; Konold, Higgins, Russell, & Khalil, 2015; Wild, 2006) suggested the importance of developing a *global view* of distributions (or, equivalently, a global view of data). Here, global view refers to the ability to investigate, recognize, and explain general, "discernible patterns in the data" (Lehrer & Schauble, 2004, p. 636). Students are supposed to recognize the location around which values are centered and the overall pattern of the shape the individual observations collectively form (Ben-Zvi & Arcavi, 2001). In brief, the hallmark of a global view is recognizing the "features not inherent to individual elements, but to the *aggregate* [emphasis added] that they comprise" (Ben-Zvi & Arcavi, 2001, p. 38). Accordingly, Lehrer and Schauble (2004) reported that students become uncomfortable when they are pushed to adopt an aggregate view of data. The researchers hypothesized that the discomfort occurs because an aggregate view of data hides of the individual data values (such as when data are represented in histograms) and shows only groups of data values. Prior to aggregation, students can point to the area in the graph that represents data about each person or observation in the data set, but this is not possible after aggregation (Lehrer & Schauble, 2004).

In contrast to the global (or equivalently, the aggregate) view of distribution (Lehrer & Schauble, 2004; Makar, 2016), a *local view* is restricted to a focus on only the individual values within the data (Bakker & Gravemeijer, 2004; Konold & Higgins, 2003). Instead of aiming to capture what a distribution collectively suggests, students continue to consider only individual observations in the distribution. For instance, students who hold a local view tend to perceive mean as a property of a particular value or observation instead of considering the mean as a characteristic of the whole distribution collectively formed (Bakker, Biehler, & Konold, 2005). Hence, these students may find it perplexing when none of the observations in a distribution has the same value as the mean of the distribution.

Distributions, through some sort of conventional graphs, provide visual representations of data in general, and variability specifically (Makar & Confrey, 2005; Noll & Shaughnessy, 2012; Peck et al., 2013). Students who accomplish the global view better understand that variability can be observed in a graphical representation than students who do not have a global view of data (Arnold & Pfannkuch, 2014; Pfannkuch & Reading, 2006). Thus, understanding how data act visually is a fundamental task that students need to carefully practice in order to understand variability more thoroughly (Franklin & Kader, 2006, p. 1).

Identifying important attributes and relationships from raw data is often challenging even for expert statisticians; hence graphical displays are used for exploring the features inherent to data. The ability to interpret distributions as represented by common graphical displays is an essential component of statistical literacy (ASA, 2005; Gal, 2002; Rumsey, 2002). Dot plots and histograms are two of the most common kinds of graphs used to depict the properties of the distribution of univariate quantitative variables (Arnold & Pfannkuch, 2014). A robust understanding of a distribution as displayed in these graphs should address shape, center, and variability as well as features such as gaps, clusters, and outliers in the context of the data (delMas, Garfield, & Ooms, 2005; Garfield & Ben-Zvi, 2005). The shape of a dot plot or histogram, for instance, implicitly indicates information about the variability of the distribution, because the shape presents the extent to which data values, in general, fall far or close to the center (Konold & Pollatsek, 2002), thus enabling one to see the pattern in a distribution (Wild, 2006). In other words, a robust consideration of variability in a distribution necessitates important observations from the shape and center of the distribution. Cobb (1999) found that middle school students' exploration of the shape of distributions could help them intuitively recognize the concepts of variability, sampling, and what data essentially suggest.

Although statistical graphs are introduced to students early at school (Friel, Curcio, & Bright, 2001), students generally fail to comprehend and use these graphical displays (Humphrey, Sharon, & Mittag, 2013). For example, students tend to limit their focus on describing individual observations (i.e., data values) in a dot plot or bins in a histogram (Konold & Higgins, 2003) or as the center and ignore other essential features in a graph (Ben-Zvi, 2004). Students may also overgeneralize the role of shape in determining variability and, for instance, may assume that a symmetric distribution should always result in less variability (Kaplan, Gabrosek, Curtiss, & Malone, 2014, p. 3). In addition, Pingel (1993) claimed that students often fail to understand that "distributions with the same mean and standard deviation can have different shapes" (p. 70). To conclude, students, even at the undergraduate level, often neglect important components that are crucial in exploiting graphical displays (Bakker & Gravemeijer, 2004; Konold & Higgins, 2003; Leavy & Middleton, 2011; Meletiou-Mavrotheris & Lee, 2010).

Students at various grade levels usually hold a limited understanding of variability, making the reasoning about variability in comparing distributions challenging (Ciancetta, 2007). Students also solely depend on their own informal ways of reasoning about variability (Jones & Scariano, 2014) even after exposure to formal statistics instruction. Thus, it is important to know some of these informal tools that students commonly employ. In the next section, I discuss students' informal ways of reasoning about variability.

Students' Informal Notions of Variability

Students usually hold a naive, cursory, or fragmented understanding of variability, which are called *informal notions of variability* in this study. These notions do not need to be incorrect; they may be locally or partially correct according to a (given) situation. In addition, they may occasionally work as helpful ways of reasoning. However, in my perspective informal notions are not robust enough overall for addressing variability thoroughly. In this section, I discuss the common informal notions students employ when reasoning about variability.

Students hold several primitive and intuitive ideas about variability (Jones & Scariano, 2014), one of which is recognizing that *things* vary. By focusing on single data values (i.e., observations) one by one in a distribution (or, equivalently, by examining values of a variable in a dataset), a student recognizes that there may be differences among observations in the dataset under investigation (Canada, 2004; Garfield et al., 2007; Jones & Scariano, 2014). English and Watson (2015) and Garfield and Ben-Zvi (2005) suggested students' recognition that *observations vary from one to another* is an essential understanding of variability especially in early school years. Similarly, Cooper and Shore (2008) found that students in early elementary grades could recognize that ideas, preferences, opinions, or qualities about the variable under investigation differ across observations. Overall, this type of understanding is particularly helpful in understanding the variability of categorical variables. Although this way of understanding variability is necessary for univariate quantitative variables, students additionally should include the measure of center and focus on how far each observation is away from the center.

In elementary grades mathematics, students collect data and visualize them using convenient methods, such as picture graphs (Franklin et al., 2007, p. 24). A statistical graph basically serves in visualizing the distribution of a variable, and students' experience with graphs helps them see the presence of variability in a distribution. It is important for students to recognize that a variable of a dataset may have a "larger" or "smaller" variability, and that there can be no variation if, for example, each observation has the same value (Garfield & Ben-Zvi, 2005).

Students' data investigations in elementary grades, which are often investigations limited to categorical and discrete quantitative variables, indicate the presence of variability. When young students talk about variability, they usually mention: consistent, similar, different, uncertain, typical, likely, more-or-less popular, common, and close to fair share. The expectation for students is to make the transition from their own colloquial understandings of variability—things that change—to more normative understandings and uses over time (Ciancetta, 2007; Kader & Jacobbe, 2013).

Elementary grade students can generate understandings that are similar to standard statistical understanding (Lehrer et al., 2011). They can reason about variability in a considerably more robust way if appropriate and substantial intervention is provided (Watson, 2009). For instance, in their study with second-grade students, Jones et al. (2001) found that students were able to reason about variability using a schema that the researchers called "close-together or far-apart"—reasoning that has a proximity to the normative statistical description for variability. Similarly, Lehrer and his colleagues' study with fourth and fifth graders in the context of measurement and data modeling (see Lehrer et al., 2011; Lehrer & Schauble, 2002) found that students were able to express variability as originating from measurement error, recognize its role in the shape of a distribution, and quantify precision of a measurement by gauging how far away each data point is from the center. Lehrer and Schauble (2004) suggested that students can develop an understanding of variability in the context of measurement and data modeling as follows:

> The context of measurement provided strong support for students' interpretation of statistical concepts related to distribution (Konold & Pollatsek, 2002). Measures of center corresponded to true scores, measures of spread to the tools and techniques employed by the measurers, and the overall shape of the distribution to the nature of error in this context. All of these qualities of distribution could be seen as emerging from students' collective activity. (p. 638)

In the following three sections, I describe the most common informal notions that are listed in statistics education studies. I name these notions as (a) has a greater range, (b) has (more) extreme values, and (c) has fewer same (or similar) values.

Has a Greater Range

When students are explicitly asked to consider variability of a distribution, they tend to equate variability with range (Ciancetta, 2007) and thus focus only on range differences among the distributions under investigation (Lann & Falk, 2003; Shaughnessy, 2007). For example, in their study with 354 first-year university students, Lann and Falk (2003) found that a greater proportion of students employed range than any other single measure of spread (i.e., MAD, IQR, and SD) when they were asked to compare the variability *intuitively* of given pairs of small raw datasets. Garfield et al. (2007) and Lann and Falk (2003) found students to be easily distracted by the differences in range values between datasets, which in turn, discourages students from investigating variability further. Lehrer and Schauble (2002) claimed that reliance on range is one of the impediments to the conceptual attainment of variability.

Has (More) Extreme Values

Students tend to put more emphasis on some of the values in a given dataset (or distribution) over other values, usually on the considerably smaller or bigger values. These values are usually called extreme values or outliers. According to Ben-Zvi and Arcavi (2001), an outlier is an individual data point (i.e., observation) that is beyond the overall pattern of a distribution. A similar description for outlier suggested by Agresti and Franklin (2015) is the "observation that falls well above or well below the overall bulk of the data" (p. 52). Although Ben-Zvi and Arcavi (2001) as well as many textbooks defined outliers as observations that fall outside of the overall trend, it should be noted that it is not precise description. Gould and Ryan (2014) support this claim by suggesting in their introductory statistics textbook that the term *outlier* has no precise definition in statistics. Gould and Ryan suggested that extreme values are outliers, thus extreme values also do not have a precise, commonly agreed-upon definition.

When asked to reason about variability, students, including most undergraduate students, appear to narrow their attention to only individual values in a distribution (Garfield et al., 2007; Shaughnessy, 2006, p. 88). In their study with middle-school teachers, Confrey and Makar (2002) found that when the research participants were asked to examine variability they limited their focus only to individual points and especially the extreme values. Noss, Pozzi, and Hoyles (1999) studied the ways pediatric nurses make sense of clinical data by focusing on average and variation. The study found that nurses pay attention only to outliers of the clinical data such as blood pressure, temperature, respiration, and pulse, perhaps because of the critical role outliers play from a clinical point of view.

Previous studies also reported that students might consciously neglect extreme values from their investigations with data (Lehrer et al., 2011, p. 732). It should be noted that handling extreme values in a dataset is usually critical in statistical investigations. Statisticians and statistically expert people flexibly handle extreme values in multiple ways such as using their expertise and according to the context or the goal of a statistical investigation. In some cases, extreme values can be evidence of an extraordinary incidence, whereas in other situations, they may be simply because of a measurement error; hence it is better to remove them from the dataset in the latter case. The key point to remember when handling extreme values is that the focus on extreme values should not undermine attending to the global characteristics of a distribution (Konold & Pollatsek, 2002; Lehrer & Schauble, 2004). To conclude, focusing only on individual data points results in failing to consider the overall characteristics of a distribution, thereby concluding an insufficient assessment of variability (Confrey & Makar, 2002; Garfield & Ben-Zvi, 2005; Lann & Falk, 2003; Peck et al., 2013, p. 25; Shaughnessy, 2006, p. 88).

Has Fewer Same (or Similar) Values

This informal notion of variability is related to students' confusion between the interpretation of variability in categorical and quantitative variables. As I discussed in previous sections, the core meaning of variability for categorical and univariate quantitative variables are different, and students usually fail to distinguish what *variability* specifically means in each case. For instance, students in Hammerman and Rubin's (2004) study focused on *heterogeneity* in their investigation of univariate quantitative variable, and they concluded that the distribution with same values repeating more often were less variable. Loosen et al. (1985) investigated 154 undergraduate students' intuitive ways of understanding variability of quantitative variables before the students began learning statistics, and the researchers found that students usually base their choice on an informal notion: how much the values differed from each other. The study clearly showed that students often fail to focus on deviations from the measure of center. Similarly, Lann and Falk (2003) found that students claimed repetitions of the same value in a dataset as indicative of less *heterogeneity*, thereby, an implication of smaller variability.

As Loosen et al. (1985) suggested, students' intuitive perceptions of variability may not be compatible with the idea behind the variability of univariate quantitative variables. According to the heterogeneity notion, a dataset with values, say, 1, 4, 4, 10, 10, 10 is more variable than a dataset with values 2, 2, 2, 15, 15, 15 because there are more different values in the first dataset (1, 4s, and 10s) in total as compared to the latter dataset, which has only two different values (2s and 15s only). Similarly, a dataset with values 10, 10, 10, 10, 60, 60, 60, 60 may seem less variable to the students than a dataset with values 30, 31, 32, 33, 33, 34, 35, 36 although the latter dataset clearly indicates a clustering around the mean value of 33. In brief, the notion lacks clustering around the center.

Loosen et al. (1985) claimed the existence of this notion to be partially due to the way variability is introduced in textbooks. According to the researchers, when textbook authors introduce variability, they usually emphasize variety among the data values as leverage but stress *the deviation from the central value* considerably less, which, in turn, delivers an incorrect signal to students about what variability means for quantitative variables.

In this section, I reviewed students' common informal notions about variability. These notions were based on *overreliance on range, a focus only on extreme values, and exploration of same (or similar) values* in a dataset. In the next section, I discuss the common student misconceptions about variability.

Misconceptions Literature on Variability

Previous studies have shown that students at various grade levels hold misunderstandings about variability (e.g., Ciancetta, 2007; delMas et al., 2005; Kaplan et al., 2014; Meletiou & Lee, 2002; Meletiou-Mavrotheris & Lee, 2010). An important aspect to note is that variability is connected to and contingent on understanding various other statistical concepts. Often, the misconceptions about variability have components that also account for the limited understanding of a concept other than the variability. Therefore, claiming a common misunderstanding as solely as a misconception about variability may potentially hinder the intricacy of the issues with students' understanding.

Graphical displays are commonly used in the discipline of statistics (Garfield & Gal, 1999); hence, they also have a significant place in teaching and learning statistics. However, elementary, secondary, and undergraduate students often demonstrate difficulties understanding and using statistical graphs (Humphrey et al., 2013). For example, K–12 mathematics and introductory statistics curricula commonly include bar graphs and histograms as graphical summaries respectively for categorical and quantitative variables (Arnold & Pfannkuch, 2014). Many students often treat bar graphs (that show category counts or percentage through bars) as if they were histograms and draw incorrect conclusions (Humphrey et al., 2013). According to delMas et al. (2005), students confuse bar graphs and histograms because of their visual similarity (each uses bars to represent data) and because the *y*-axis in both graphs represents the frequencies of each bar (bars in bar graphs and bins in histograms). To be able to distinguish bar graphs and histograms from each other, students need to recognize that bar graphs are used to display the counts or percentages of categorical data (Bock et al., 2012, p. 67), whereas histograms are used to visualize the global characteristics of univariate quantitative data (Arnold & Pfannkuch, 2014; Humphrey et al., 2013, p. 72).

Students should also understand that reasoning about variability based on these commonly used graphs is different. When investigating for greater variability in histograms, students are expected to choose the one that has the least central clustering. Clustering more closely around the mean indicates less spread from the center, therefore a relatively smaller SD (delMas & Liu, 2007, p. 112). However, clustering around the center is inapplicable in bar graphs because a typical center value does not exist in categorical data (unless there are two categories and the proportion of success or failure could be regarded as a measure of center). The heuristic for bar graphs is that "the closer the distribution is to the uniform distribution (all categories having the same relative frequency) in a bar graph, the more variability there is in the data" (Kader & Jacobbe, 2013, p. 18).

Problematically, students often fail to coordinate the information presented in the frequency axis (i.e., the *y*-axis) and the variable values axis (i.e., the *x*-axis) of histograms (Cooper & Shore, 2008, p. 7). In other words, students confuse the information the vertical and horizontal axes of a histogram provide about variability. The horizontal spread provides the primary source of information about variability; thus students should recognize that the key aspect to consider in a histogram is the *horizontal* dispersion (Gould & Ryan, 2014, p. 83; Vermette & Gattuso, 2014, p. 3). Although the horizontal spread indicates some information about variability of a distribution by displaying the deviation from the central tendency, students often tend to mistakenly think that *vertical* dispersion could be used directly to reason about variability. For instance, students often claim that presence of clusters of bars of different size in a histogram implies larger variability. In other words, they claim that an approximately uniform distribution would indicate less variability because of the similar frequencies. This lack of understanding is typically conveyed in students' expressions such as, "It is most evenly (or uniformly) spread out, thus less variable." In relation to students' reckless use of vertical dispersion, many students assume that existence of more "bumps" in a histogram indicates that the distribution is more variable (Cooper & Shore, 2008; delMas & Liu, 2005, 2007; delMas, Garfield, Ooms, & Chance, 2007; Kader & Jacobbe, 2013; Meletiou & Lee, 2002). Students think that a histogram with a larger number of different *bin values*—which are basically intervals of the value of a quantitative variable—should imply a larger standard deviation (delMas et al., 2007). Interviews and large-scale surveys conducted by Meletiou and Lee (2002) with college students provided evidence that students think that a bumpier histogram should indicate larger variability even though such histograms imply the inverse in reality. In sum, students often fail to see that larger differences in frequencies (heights) of the bins of histograms does not directly relate to a larger variability (Garfield et al., 2007).

A common source of students' misunderstandings is applying a rule, property, or concept beyond "its range of legitimate applicability" (diSessa, 1993, p. 116). An example of this type is the use of the information that the shape of a graph could provide beyond its applicability. For instance, some students assume that a bell-shaped distribution should always imply less variability (Kader & Jacobbe, 2013). It is commonly true that a bell-shaped distribution indicates a relatively smaller variability when compared with distributions that are not bell-shaped. This is not always the case, however; a uniform distribution with a smaller range, for example, may be less variable than a normal distribution with a larger range. In addition, a distribution with all the same values indicates no variability. In sum, students should be careful when generalizing the information that the shape of a histogram can convey about variability.

In this section, I have discussed the most commonly reported variability misconceptions. In the following section, I briefly discuss the work of Garfield et al. (2007) and Lann and Falk (2003) because they provide a relevant foundation for my research problem. Based on the analysis of literature, I then suggest the direction that research should lead in order to provide more insight on students' reasoning about variability.

Reasoning about Variability: What Is Missing?

Research studies on students' reasoning about variability have suggested important conclusions. In one of those studies, Garfield et al. (2007) investigated the development of undergraduate students' understanding of variability in an upper division statistics course, and examined students' understanding of the concept with pre- and post-instruction assessments. The researchers found that given a distribution, students tended to rely on the idea of range and individual observations to reason about variability. The researchers also suggested two informal notions that are foundational in students' understanding of variability: "Variability is represented by overall spread and differences in data values (e.g., not all values are the same)" (p. 142). The researchers concluded that students often fail to focus on where most of the data in a distribution are located relative to its center.

Learning variability is challenging for students at all grade levels. In order to address the problem, Garfield et al. (2007, p. 142) proposed a hypothetical learning trajectory for developing an understanding of variability. According to the trajectory, statistics instruction should begin with addressing the basic recognition that data vary. Students then should be directed to explore some of the possible sources of variability, such as measurement, natural, and chance variability. Next, the instruction should enable students to investigate commonly used graphical representations in order to compare variability among different datasets. This step could also help students examine the effects of bumps and clumps (in a histogram) on variability and begin to recognize the importance of overall spread and clustering in variability. The instruction then should focus on building connections between measures of center and measures of variability. Finally, students should be able to understand the characteristics of the measures of variability, such as whether or not a measure is resistant to extreme values (e.g., range is resistant, whereas SD is not resistant, to extreme values). Understanding these characteristics can help students cultivate the ability to select the optimal measure of variability in particular situations.

Although Garfield et al. (2007) claimed the steps mentioned above as a hypothetical learning trajectory, their work more closely resembles a curricular plan that one could follow in teaching variability. In other words, these steps essentially focused on the work related to teachers, but did not provides insights on how students could potentially learn the concept. In addition, the researchers were not able to provide research-based evidence or a well-framed rationale for their “trajectory,” and they did not implement the trajectory to verify whether folllowing the suggested steps substantially improves students’ understanding of variability.

Although more work is needed, the steps that Garfield and her colleagues proposed may be helpful in investigating students' informal notions about variability.

Lann and Falk (2003) explored first-year university students' selection of measures of variability by asking students to intuitively choose the dataset that had more *heterogeneity* and *inequality.* The researchers consciously avoided using expressions that could hint at the idea of clustering about the mean, gap, or extreme values and technical terms, such as *variance, SD,* and *range*. The study suggested that a greater proportion of students used range (without explicitly using the term in their responses) more than any other single measure of spread (e.g., MAD, IQR, or SD). However, Lann and Falk's study presented two problematic aspects. First, the researchers intentionally attempted not to use standard terminology, such as *range*; as a result, students also did not explicitly say that they had used the idea of range in assessing heterogeneity and inequality in given datasets. Instead, the researchers inferred the results based on the fact that students more commonly selected the datasets that in fact had a larger range. Next, asking students to explore heterogeneity and inequality in order to investigate students' intuitive judgment of variability for univariate quantitative variables does not seem to be an appropriate approach. As Loosen et al. (1985) pointed out, seeking heterogeneity in univariate quantitative data provides limited information about the quantitative dispersion of a dataset. As discussed when introducing variability for different types of data, variability for categorical data is summarized as *how many,* variability for quantitative data is summarized as *how much,* and heterogeneity refers to the *how many* approach. Overall, although Lann and Falk (2003) overlooked these problems in their study, the instruments used in the study and findings of the study were useful in my investigation of students' reasoning about variability.

As mentioned before, studies have shown that when asked to reason about variability, students often use such informal notions as reliance on range, individual values (especially the extreme ones), and the degree to which data values are similar to each other. These informal notions could contribute to students' development of a more robust understanding of variability. An important question to investigate then is to what extent could these notions be exploited in students' development of the normative meaning of variability. If a particular type of informal reasoning has potential to develop an understanding of variability, then it warrants effort to uncover and investigate the ways in which students might benefit from the notion as they learn variability.

Students often fail to follow standard ways of reasoning; instead, they follow strategies that are more meaningful or appropriate for their intuitions. My review of the literature suggests that previous research studies have not closely examined students' informal notions. For example, there has been no research study setting in which students could not use *range*—because the given datasets to be compared had equal ranges—when exploring variability. Investigating students' reasoning when datasets have equal ranges but have different distributions may help researchers diagnose other possible reasoning mechanisms that students use and whether those mechanisms support a more standard understanding of variability. In addition, little is known about how students reorganize their reasoning when their informal notions remain ineffectual.

Although previous studies yielded important results about students' particular ways of reasoning about variability, much work remains to be done in the study of students' conceptualization of variability. Previous studies failed to consider to what extent the particularities of a dataset or distribution are influential in students' reasoning. In other words, researchers often did not address the relationship between students' reasoning and the prevalent characteristics of distributions that students were asked to reason. For example, a student might have a heuristic that if a distribution has extreme values then the distribution should be more variable; but what if the student is asked to reason about the same distribution with the extreme values removed? Similarly, the literature fails to provide enough evidence to conclude that the only reason for students to rely on range is simply because it is easy to use. Therefore, the question that merits further investigation is how do students reason about variability in situations in which the datasets to be compared (a) have equal ranges; (b) have no extreme values; and (c) have the same (or similar) number of repeating values.

ASA (2005) and Franklin et al. (2007) highlighted the overall role of context in statistics keeping it "alive" throughout any statistical investigation. Much of students' intuitive knowledge consists of loosely connected pieces of knowledge, the activation of which is highly dependent on context (Elby, 2000). Meletiou and Lee (2002, p. 33) claimed that students' reasoning about variability is heavily reliant upon the type and context of statistical activities, materials, and tasks. My review of the literature, however, found no study that delineated the role of context on students' understanding of variability. Therefore, another necessary question to seek answers is the role of context in terms of students' ways of reasoning about variability.

Statistics education research should investigate how to promote student-generated understanding in the classroom (Lehrer & Schauble, 2004, p. 670; Leinhardt & Larreamendy-Joerns, 2007; Shaughnessy, 2007). Few research studies envision how students could develop a formal understanding of variability based on their own informal notions. Hence, informal notions about variability warrant efforts to uncover and investigate how they are mediated as students are exposed to instruction on statistics. For example, a formal introduction to measures of center (e.g., mean, median) and measures of spread (e.g., SD) could build on students' intuitive notions of center and variability (Jones & Scariano, 2014; Shaughnessy, 2006, p. 94). We should learn more about how students use and revise their informal notions when they encounter different situations and, in particular, when more standard meanings and measures of variability seem more appropriate for students to use.

By exploring the ways students' reasoning associates with particular aspects of the distribution or data (on which they are asked to work), a research study may shed light on undergraduate students' reasoning about variability. Such a study has the potential to inform the extent to which undergraduate students are able to reason about variability when the informal notions they hold are not applicable. The study may also provide more information on possible learning trajectories for understanding variability. In addition, such a study has the potential to offer more information about the role of context in students' ways of reasoning.

Theoretical Framework

I begin this section with a clarification of terminology, specifically, the use of the terms *context* and *situation*. In this study, *context* refers to the real-life circumstance or a scenario presented in a statistical question. *Situation* pertains to particular aspects of a statistical question (e.g., whether raw data or a conventional graph is provided for exploration) or characteristics of a distribution (e.g., whether the datasets to be compared have equal ranges).

Additionally, I define informal notions of variability as pieces of knowledge that students construct and reconsider as they encounter any new data, graphical display, or context. Students' informal notions can (a) be regarded as loosely connected pieces of knowledge that play a role in their conceptualization of variability, (b) be valuable and convenient tools for students to use depending on a *particular* question, dataset, or distribution, but (c) have definite limits in application. Informal notions might have a disparate impact depending on the situation. In other words, although some informal notions dominate in particular situations, they may become misleading or insufficiently rigorous in others. As a result, students may apply to their informal notions in different ways and weights.

Considering the goals of the present study, I developed the theoretical framework for this study using diSessa's (1988, 1993) *knowledge-in-pieces* (KiP) *epistemological perspective* and its core components. In the following, I explain my understanding of the perspective and the way I adapted it for this research study.

Knowledge-in-Pieces Epistemological Perspective

Andrea diSessa (1993) proposed the knowledge-in-pieces epistemological perspective to explain students' development of knowledge in the learning and understanding of Newtonian physics. According to diSessa, actual knowledge elements are more diverse and smaller than a typical textbook presentation would suggest. In addition, although knowledge elements may be *loosely* connected to each other, they are *strongly* tied to particular situations so that novices can make sense of these elements. Conceptual change and knowledge growth—useful terms to study novice-expert knowledge—can occur by constructing new knowledge elements, coordinating and reorganizing both emerging and prior elements, and extending and constraining the use of particular elements of knowledge according to certain situations (Izsák, 2005, p. 5).

It is important to note that diSessa (1993) used the terms *knowledge elements* and *knowledge resources* to explain his theory of knowledge growth, but not as terms or constructs with specific meaning in the knowledge-in-pieces epistemological perspective. Moreover, diSessa also used the terms *novice* and *expert* with their common meanings. The term *novice* usually refers to a student or a person who is new to a subject or concept, and *expert* refers to a scholar or an experienced professional within a field who holds a profound understanding of that subject or concept. According to this description, undergraduate students beginning to learn statistics can be regarded as novices of the statistics subject.

Core components of KiP. One of the core constructs of KiP is *phenomenological primitives (p-prims)*, described by diSessa and Sherin (1998) as the "explicit treatment of representation, origins, and development of intuitive knowledge" (p. 1187). P-prims are elementary knowledge structures that novices abstract from their experience in order to explain and give meaning to a phenomenon and justify their decisions (Wagner, 2006). According to diSessa (1988, 1993), p-prims are often abstracted from common experiences, and they are relatively primitive. DiSessa (1993, p. 112) also posited that p-prims usually need no further justification from the novice's point of view, and consequently, novices will not seek further clarification for their p-prims.

DiSessa and Sherin (1998) claimed that different p-prims may be evoked in different situations. Similarly, Wagner (2006) suggested that novices may utilize different combinations of p-prims or other knowledge resources to make interpretations in different situations or for different aspects of a particular situation. In a case of conflict, novices usually have no mechanisms that lead them to decide which p-prim should be applied (diSessa, 1993, p. 114). Also, a hierarchy among different p-prims starts to form as students gain more experience and move toward expertise.

These ideas align with the way I defined informal notions of variability for the present study. Similar to diSessa (1993), I claim that, although students may hold a number of informal notions, these notions tend to activate "in appropriate circumstances" (p. 112), such as when a particular characteristic is highlighted in a distribution. For example, if a student is given raw, unordered data with some data values repeating (i.e., having the same numerical values more than once in a distribution), the student may interpret variability as *difference* and compare variability across distributions according to the difference notion. In this specific case, the student may find comparing ranges across datasets to be inconvenient and inappropriate. The same novice student, however, might also decide to use range if the observations in datasets to be compared are ranked in order, assuming that reporting a comparison of ranges across datasets sufficiently addresses variability. One possible motivation for employing range is that ordering data values foregrounds the difference between range values; thus, using range may seem convenient from the student's perspective. In conclusion, as students develop an understanding of variability, they should begin to see that one or two characteristics (Wagner, 2006) of a distribution, such as having a larger range or extreme values, might be inadequate to support the claim that the distribution has a larger variability.

Another core component of KiP is *coordination classes*. DiSessa and Sherin (1998) proposed coordination classes as complex sets of methods and strategies novices use to gather substantially more information from their observations and experiences. A coordination class is the competence of an individual to see a particular class of information in the world. According to diSessa and Sherin (1998, p. 1185), learners use coordination classes to understand how scientific concepts function in forming explanations and problem solving. Overall, the purpose of a coordination class is to specify the nature of knowledge structures (*structures* here being used, as is common in knowledge-in-pieces literature, in its ordinary sense) that are assumed to underlie complex conceptual understanding.

Coordination classes have two primary structural components: *readout strategies* and the *causal net* (diSessa & Sherin, 1998). The first component, *readout strategies,* deals with determining “how characteristic attributes of a concept are attended to or seen in a given situation” (Wagner, 2006, p. 7). According to diSessa and Sherin (1998), different problems require using a combination of notions, and multiple readout strategies may be needed to reason about a concept. The other structural component of coordination classes, the *causal net*, is “the set of inferences that lead from observable information to the determination of things that may not be directly or easily observable” (diSessa & Sherin, 1998, p. 1174). An individual’s causal net is the “general class of knowledge and reasoning strategies that determine when and how some observations are related to the information at issue” (diSessa & Sherin, 1998, p. 1176).

The expectation is that the particular aspects of a person’s readout strategies and the causal net will be unique to that person. This expectation suggests that novices (which would include most students) depend primarily on their p-prims, which could provide different coordination strategies for novices when they are compared to the coordination strategies that experts use. Considering p-prims and coordination classes together, diSessa and Sherin (1998) claimed that p-prims are “too small and isolated to constitute a coordination class” (p. 1179). The existence of a coordination class in someone’s reasoning implies that various p-prims play a role in his or her causal net.

Coordination classes are more applicable to more robust and relatively stable student understanding over an extended period of time (diSessa, Sherin, & Levin, 2016). Because the present study was more focused on relatively brief segments of students reasoning but not on the complex conceptual understanding, I did not attempt to incorporate coordination classes into the study.

KiP in Investigating Reasoning about Variability

Previous research studies suggested that students, who are generally considered to be novices, reason about variability incoherently and inconsistently across different situations. The knowledge-in-pieces epistemological perspective supports the observation and delineation of these types of situations (Jacobson & Izsák, 2014, p. 49). I predicted that the use of diSessa's perspective for my study could be helpful in focusing on the knowledge resources students could depend on as they make judgments about variability.

Wagner (2006) suggested that different *situations* have their own affordances and facilities with respect to students' reasoning about a concept. Similarly, Meletiou and Lee (2002) claimed that students' reasoning about variability is heavily reliant upon both the particularities of the task explored, which refers to situations and the contexts in which the tasks are situated. For instance, undergraduate students usually find working with raw data (even for small datasets) more difficult than working with graphical displays. Use of knowledge-in-pieces perspective may be instrumental to investigating ways in which the characteristics of distribution, graphical representation, and statistical context cue different ways of reasoning (Wagner, 2006). Moreover, Gould and Ryan (2014) advised in their introductory statistics textbook that the choice of most appropriate measure of variability to use in describing a distribution should be based, in part, on the shape of a distribution. Accordingly, I expect students to be more inclined to use the idea of central clustering and standard deviation in situations modeled by a normal distribution and to fail to consider these ideas when the distribution of data has a different shape.

The research studies mentioned in the present chapter agreed that students' reasoning about variability is lacking in multiple ways. In addition, students may switch to a different way of reasoning depending on the type of data and graphical representation that are given (Perry &

Kader, 2005). As a result, comparing and contrasting ways in which students recruit their informal notions of variability when they reason about graphical displays with different characteristics can provide new insights about students' statistical understanding (Perry & Kader, 2005). Characteristics of the problem, situation, or contexts are important in students' reasoning. Therefore, I considered these premises when creating tasks for students to work on for this study. Overall, by attending to the forms and types of knowledge crucial in knowledge growth, the knowledge-in-pieces epistemological perspective has the potential to provide a fine-grained way to investigate students' informal notions of variability.

It is necessary to investigate how students reorganize their reasoning when elements of their reasoning about variability are inconsistent with and in direct contradiction of each other. *Cognitive conflict,* which was coined by Strike and Posner (1992), occurs when one recognizes that his or her experience or informal notions are not consistent with his or her other notions. Experiencing such a situation may yield opportunities for students to refine their existing conceptions. For instance, it merits investigating, from a knowledge-in-pieces epistemological perspective, how students reason about variability if using range and focusing only on extreme values yield conflicting conclusions.

One potential drawback of the use of the knowledge-in-pieces epistemological perspective is relates to its assumption that physical observations and prior experience constitute an important part of novice learning. On the one hand, one can claim that the assumption could be invalid when learning statistics in general and understanding the notion of variability specifically are considered. In other words, students may have little scholarly experience with statistics and with the concept of variability prior to taking a statistics course, making the comprehension of fundamental statistical ideas inaccessible. On the other hand, it is also reasonable to think that although students may not have much experience with scholarly statistics, they may still accumulate experience, because they are constantly exposed to data and graphical displays everywhere along with various clues to variability. Therefore, students may unconsciously try to coordinate their experiences with statistics in reasonable ways. As a result, whether or not statistics learning and reasoning about variability can be rigorously scrutinized with a knowledge-in-pieces epistemological perspective is an open question that was investigated in this study.

CHAPTER 3

METHODS

In this chapter, first, I describe the research setting and context for the study by providing information about the introductory statistics course and how the course is run. Next, I introduce the research participants of the study, as well as the recruitment process followed for the interviews. Then, I explain the data sources and the data collection procedures. The chapter ends with a description of the analysis of data.

Research Setting

The data for the study came from students taking STAT 2000, a multi-section, 4-credit, introductory statistics course offered each spring, fall, and summer semester at the University of Georgia (UGA). The university bulletin describes the outline of the course as "the collection of data, descriptive statistics, probability, and inference. Topics include sampling methods, experiments, numerical and graphical descriptive methods, correlation and regression, contingency tables, probability concepts and distributions, confidence intervals, and hypothesis testing for means and proportions" (UGA, 2015).

During the time of the study, there were approximately 1,200 students enrolled in the various sections of the course. The course had both lecture and computer lab components. Students selected their lecture and the computer lab sections when registering for the course. The weekly 150-minute lecture component was divided evenly into two class sessions (on Tuesday and Thursday) or three (on Monday, Wednesday, and Friday). Lecturers and teaching assistants who were statistics graduate students (graduate teaching assistants, or GTAs) taught the sections of the course in a lecture format to large classes with approximately 200 students in each section. Teaching assistants (TAs), who were also statistics graduate students, delivered the weekly 50-minute computer lab component. TAs usually taught several sections of these computer labs. Overall, approximately 4–6 lecturers and GTAs taught the lecture component, and 10–12 TAs taught the computer lab component of the course. A coordinator from the Department of Statistics oversaw the course. I henceforth refer to GTAs, lecturers, and TAs as teaching personnel, without maintaining a distinction among them because such a distinction was not crucial for this study.

The course included several assessments assigned throughout a semester. These assessments included five tests (the last one being optional), twenty homework assignments, and ten computer lab assignments. Students could complete the computer lab assignments during their assigned lab hour. The tests constituted 80% of the final grade of the course, and the remaining 20% was distributed between homework and computer lab assignments in differing weights (for more information, see Jennings, 2014; UGA, 2015).

Students were required to use a course assignment platform called WebAssign (https://webassign.net/) to take the tests, to access computer lab assignments and homework, and to input their responses. Tests, homework, and computer lab questions either were in multiple-choice format or asked for a single numerical answer. Students received slightly different assignments from each other, which was achieved by randomizing (a) the numbers used in the questions, (b) the order in which the questions appeared in the tests, and (c) the order in which answer choices appeared on multiple-choice questions. The WebAssign provided immediate feedback for homework and computer lab assignments. It allowed three attempts for each question before a final credit was awarded to a response. On tests, however, students were limited to one attempt for each question. The WebAssign performed the grading automatically; no human scoring existed in the then-current delivery of the course.

Participants

The population of interest in this study was undergraduate students who enrolled in the introductory statistics course during the spring, summer, and fall 2016 semesters. Students who registered for the course usually had different backgrounds and profiles, for example, in terms of the number of years in college, intended majors, quantitative reasoning courses taken, and prior experience with statistics.

Students' responses to homework assignments were collected from all of the enrolled students in the course who submitted their assignments during the Spring 2016 semester. For the second source of data, a series of two to three interviews were conducted with students who were enrolled in the course. Six students were recruited for interviews, but only four of those students' interviews were included in the analyses. Interview studies typically yield a large amount of data; thus, keeping the number of interview participants to four appeared to be an appropriate choice in order to engage in in-depth analysis.

For the recruitment of interview participants, I first contacted the TAs of the course. My request for interview participants included the recruitment letter (Appendix A) in which I explained the research study and what was expected from interview participants. The inclusion or exclusion of the students for a subsequent interview was based on their performance in the first interview and their openness to thinking aloud. Fulfilling any of the criteria was challenging, especially for recruiting students for the first interview, because there were limited opportunities to become acquainted with the students before the first interviews were conducted. Therefore, recruiting more students for the first interview seemed to be an appropriate strategy to meet the selection criteria and assure that the data were rich enough to investigate the research questions.

I started the interviews with two students in Summer 2016. One of those students was not able to give extended answers to the interview questions and explain his way of thinking. As a result, I did not invite him to the second interview. The other interviewed student was able to work on the interview tasks and explain his reasoning, so I continued to interview him. I was able to ask him all of the interview tasks in three 1-hour interviews. Then, I started to conduct interviews with four students in Fall 2016, and was able to continue interviewing these four students during the semester. Although the second interview participant I interviewed during Summer 2016 provided extended answers, I decided not to include his data in the analysis for two reasons: (a) the other four interview participants were enrolled in the introductory statistics course in the same semester except this student, and (b) including the student's interview data did not seem to substantially increase the overall richness of the data that I expected to have by including the other four students' data. Overall, I decided to use four of the students' interview data from the same academic semester in the final analysis for their richness and clarity in demonstrating the construct of interest for the study. Table 1 provides background information about these students (real names replaced with pseudonyms) who were female and all in their second year at UGA.

Table 1

Interview Participants

Name	Major	Statistical Experience
Ocean	Biology, Pre-dentistry	Took an AP Statistics course in high school but not the exam
Karen	International Affairs, minor in Women's Studies	Took an AP Statistics course at high school
Chloe	Political Science, minor in Philosophy	Took an AP Statistics course
Britney	Exercise and Sport Science	No experience

Data Collection

Utilizing multiple data sources gives access to forms of information that are more effective in addressing research questions than single sources would be (Maxwell, 2013). The first data resource was in the form of student responses to homework questions in the introductory statistics course during the Spring 2016 semester. The data consisted of student responses to a combination of eight open-ended and multiple-choice homework questions collected through WebAssign. The second source of data was the video recordings of the interviews and written artifacts that the interview participants produced during the interviews. I individually interviewed the selected students during the summer and fall semesters of 2016. In the following sections, I explain the details of both data sources and the data collection procedures.

Homework Assignments

The first stage of the data collection involved gathering student responses to eight multiple-choice and open-ended questions embedded in the online homework assignments (see

Appendix B). The data were collected from all of the students in the statistics course who submitted their responses to the assignments in the Spring 2016 semester. The approximate number of students enrolled in the course was 1200, and usually the majority of the students submitted responses to the assignments.

Among the eight questions in total, Questions 1, 3, and 7 (as listed in Appendix B), were originally created and piloted by a statistics education research team led by Jennifer J. Kaplan, who is a statistics education researcher working at the same university. These three questions were included in the study because they essentially asked students to compare variability of distributions as visualized in various dot plots and histograms. The distribution of data in these dot plots and histograms limited the use of some of the aforementioned informal notions of variability. Hence, it was my expectation that responses to those questions had potential to shed light on the first research question of the study.

In order to have more questions that could be helpful in answering the research questions, I generated new questions. For new questions, first, I examined the statistics education literature (especially dissertations, research and practitioner journal articles); research project materials such as LOCUS (Jacobbe, Case, Whitaker, & Foti, 2014), ARTIST and CAOS (delMas et al., 2007); college introductory statistics textbooks (such as the *Intro Stats* by Bock et al., 2012; *Introductory Statistics: Exploring the World through Data* by Gould & Ryan, 2014); and curricular and report type documents (such as ASA, 2005; Franklin et al, 2007; Franklin et al., 2015). Then, I located pertinent materials such as datasets, graphs, questions, or contexts and modified them to write new questions. The statistics education research team at UGA inspected these new questions and suggested modifications as necessary. I revised the questions further to give them their final form. Then, I included these questions in the online homework assignments.

Table 2 shows how each question targets the research questions of the study and whether or not without-context version exists for a question.

Table 2

Overview of Homework Questions

Item	Research questions addressed	Had also a version without context	Had a multiple-choice component
1	General	No	No
2	1.a	Yes	Yes
3	1.a & 1.b	No	No
4	1.b	No	No
5	1.a & 1.c	Yes	No
6	1.a & 1.b	No	Yes
7	1.a & 1.c	No	Yes
8	1.c	No	Yes

Note. Question 1 did not specifically address any of the research questions of this study but was included for the reason provided in Appendix B.

These research-related questions were different from the rest of the questions of the homework assignments for the STAT 2000 course in three ways. First, each of these questions involved an open-ended component in which students were asked to explain or justify their choice. Second, in contrast to the three attempts provided for students to resubmit their responses in typical homework questions, students had only one attempt on these research-related questions. Finally, students earned points automatically upon uploading their responses to these questions irrespective of the correctness of their answers.

Interviews

In-depth access to student reasoning is essentially more possible in an interview environment because of the dynamic and interactive nature of interviews (Maxwell, 2013). Hence, I incorporated interviews into the research study by recruiting students for a series of "task-based interviews" (Goldin, 1997, 2000). In these interviews, students were asked to think aloud while working on statistical tasks.

I decided to use the same questions in both homework assignments and interview tasks so that the findings from the data gathered in one could corroborate those of data gathered in the other. Therefore, the development of the interview protocols was based on the eight questions that were employed as homework questions. In brief, I was able to develop interview tasks through modifying the homework questions by considering the privileges and opportunities that a typical task-based interview environment could potentially provide. For example, although students were presented with one of two versions of a question, such as the one with (or without) a context in a homework assignment, both versions could be asked in an interview. In addition, interviews provide opportunities to ask multiple follow-up questions based on the type of responses students elaborate during the interviews. The final form of the interview protocols is in Appendix C.

The first interview with each participant usually started with a short conversation about the educational background of the interview participant, such as major and the number of years spent in college. Students who take the introductory statistics course could also have different prior experience with statistics. In order to gain insight into that aspect, I also asked the interviewees to talk about their experience with statistics at the high school and college levels.

After this short conversation, I asked each interviewee to describe variability and suggest terms and words that they could use to explain variability (see Appendix C). Next, I asked the interviewee to start working on the statistical tasks. I explained that the main goal of the interviews was to have students provide as detailed an explanation of their reasoning as possible. I reiterated during the interviews that I was more interested in the interview participants' ways of reasoning rather than whether or not their responses were correct. I asked clarifying and follow-up questions and encouraged the participants to talk freely. During the interviews, I actively questioned each participant's reasoning by pointing out the apparent inconsistencies in her approach, but I refrained from taking an evaluative role. Each interview lasted approximately an hour, was video recorded with a single camera pointed at the paper on which the interview participant put her work, and were audio-recorded with a second voice-recording device.

Data Analysis

In this section, first I explain the data analysis method for the video recordings of the interviews. Next, I describe the analysis of student responses to homework questions. I follow this order also in reporting the results in the next chapter.

Analysis of Interviews

The mathematics education literature includes many methodological approaches to the analysis of the video data. These approaches commonly follow verbatim transcripts, line-by-line analysis, and less direct use of video recordings. In the following paragraphs, I explain the analysis method I employed in this study by first providing a rationale for it.

It is common to rely on transcripts in analyzing video data. Powell, Francisco, and Maher (2003), however, asserted that relying on transcripts in analysis "makes it difficult to keep contact with one's theoretical perspective while sampling" (p. 411); thus Powell et al. suggested working directly from the video recordings instead of verbatim transcripts. The researchers provided a seven-step analytical model to analyze video data. The phases are (a) viewing the video attentively, (b) describing the video, (c) identifying critical instances, (d) transcribing (selectively), (e) coding, (f) building a storyline of the issue under investigation, and finally (g) composing narrative. Powell et al. claimed these phases to be “interacting” and “non-linear” (p. 413). In the next paragraphs, I explain how I adapted Powell and colleagues’ model to my data analysis.

For the analysis of the interviews, I focused on capturing aspects underpinning participants’ informal notions of variability in relation to the research questions and the theoretical framework for this study. As the first step to the analysis, I viewed the videos to become familiar with the approaches the participants used when working on the interview tasks. Findings of the previous studies and examples I discussed when introducing the theoretical framework were indicative to some extent of what to expect from interviewees’ responses.

For the second step of the analysis, I wrote descriptive portraits with time stamps for reference. This step included my description of how the interview participant approached the tasks and responded to follow-up questions. In brief, in this step I created a running summary of each interviewee’s ways of reasoning about variability, which included direct quotes if necessary. This step provided an effective way to deal with interview data compared to the difficulty of managing the high volume of information video recordings could potentially present. According to Powell et al. (2003, p. 416), it was important to note that a written description created in this step should be more descriptive in nature and not interpretive in order to be open to other possible interpretations of the data in the next steps.

Powell et al. (2003) suggested that it is crucial to transcribe the parts of the video recordings appearing to offer the richest information about the research question of a study. Hence, the next step was identifying *critical instances* of students' reasoning in the interviews. As Wagner (2006) suggested, the focus was on the type of *knowledge resources* the interviewees depended on when reasoning about variability. For instance, an aspect of knowledge pieces on which interview participants relied such as an explicit attention to extreme values on a given dataset might be counted as a *critical instance*. In addition, any tension that arose because an interviewee's different ways of informal reasoning conflicted with each other (e.g., across different tasks of the same interview and across the interviews) might be regarded as a critical instance. Identifying these instances was important because they provided insights into students' reasoning that has not been documented in the literature. Aspects of problem situations or tasks across different problem types and contexts that led students to focus on central clustering were also counted as critical instances.

Makar and Confrey (2005) argued that students may articulate their understanding of variability by using nonstandard language. Hence, students' use of language was regarded as indicative of their informal ways of reasoning about variability. As Makar and Confrey (2005) and Ciancetta (2007) suggested, students may use expressions such as *evenly distributed, bulk of the data, bunched up, clustered*, *spread out,* or *overall spread* to talk about variability. Although use of these words indicates evidence for addressing the concept, they are not indicative of whether or not students take into account a measure of center. Use of expressions such as spread from *the center* or clustering to *the center,* however, more clearly indicates the consideration of the center of a distribution in reasoning about variability. In brief, robust reasoning about variability should include students' use of language that includes the idea of clustering around the center.

Detecting some of the patterns and inconsistencies of reasoning and determining how the interview participants reframed their ways of reasoning as they worked on different questions, datasets, and distributions were crucial. In addition, I attempted to focus on participants' repeated patterns of argument and interpretations that were not consistent with normative statistical reasoning. I was particularly interested in the relationships between participants' approaches to tasks and the characteristics of the tasks. This step also included writing summaries that captured each interviewee's main ways of reasoning. I used these summaries to identify reasoning strategies that interviewees used in the subsequent interviews.

The interview participants' particular ways of reasoning about variability were noted, and related parts of the videos recordings were transcribed. Next, the coding of the video segments took place. In addition, critical events were transcribed and coded in order to achieve a more detailed review of students' reasoning. The codes were refined based on a repeated review of the video recordings.

The next two steps of the analysis, *building a storyline* and *composing a narrative,* took place after the coding procedure. Some of the transcribed events were kept as episodes in order to exemplify students' reasoning in the results sections. I built a story line for each research question of the study. As Makar (2016, p. 9) suggested, the whole interview video recording may not be possible to report. Instead, a coherent story could enable the reader to follow the theoretical ideas presented through interpretations of what is observed in interviews.

Crucial issues to remember when writing in this phase concerned sources of an interviewee's ways of reasoning, evidence and counterevidence, and alternative interpretations one could make based on the same data. As a result, I looked for additional data that might support, undermine, or improve my interpretations. In addition, I discussed a small part of my preliminary results with a capable colleague (who was also studying students' statistical reasoning) and asked him to suggest counterarguments to or alternative interpretations of my findings.

Analysis of Homework Questions

The analysis of student responses to homework questions was based on investigating the major reasoning strategies and the notions students presented in their responses. Overall, I attempted to examine student responses for each question in order to explore the extent to which common types of reasoning existed in the larger student population.

For this purpose, I aimed to code at least 100 randomly selected student responses for each question using Arnold's (2013) distribution framework, which I introduce in the following paragraph. Because students addressed variability less frequently than expected in their responses, I needed to code more than 100 responses for some of the homework questions.

Arnold (2013) developed a framework for analyzing students' descriptions of distributions. According to the framework, features of a distribution can be organized under five overarching statistical concepts; (a) contextual knowledge, (b) distributional, (c) graph comprehension, (d) variability, and (e) signal and noise. Arnold enumerated 28 specific features of distributions in her study, and each of these features, such as overall shape, modal groups, or whether gaps or outliers exist, could be categorized as belonging to one of the five main concepts. Because I focused on students' reasoning about variability in this study, I employed the features that Arnold proposed specifically for the overarching statistical concept of variability in her framework. Table 3 presents the variability dimension of Arnold's distribution framework.

Table 3

Variability in the Distribution Framework in Arnold (2013)

Overarching statistical concept	Characteristics of distribution	Specific features measures/depictions/descriptors
Variability	Spread	16. Range, 17. IQR 18. Range as an interval 19. Interval for high and/or low values 20. Interval for groups
	Density	21. Clustering density 22. Majority (mostly, many) 23. Relative frequency

I started to code student responses based on the features given in the framework. After the first trial of coding was completed, I often needed to collapse similar features into one category, such as giving the same code to the responses that mentioned either range (code 16) or IQR (code 17) because there were only a few instances of each across the large set of student responses. At other times, I needed additional categories in order to code responses such as the ones that discussed the role of extreme values on variability. Overall, Arnold's framework served as a first step in the analysis of student responses, but its influence was less prevalent in the subsequent iterations of coding. The framework was especially useful in coding the responses to the first, fourth, sixth, and seventh questions. For the remainder of the questions, the categories in Arnold's framework were considered, but the coding was held open to the new codes, especially because the characteristics of the distributions in these questions could trigger different ideas for students to explain their reasoning. In addition, as Makar and Confrey (2005) claimed, students often articulate their reasoning using nonstandard language. Accordingly, I did not limit the coding procedure to Arnold's (2013) categories. Overall, I repeated the coding procedure for each question at least twice for the reasons I discuss here and to increase the reliability of my coding. Table 4 shows the total response given by students and the number of coded responses for each question.

Table 4

Summary of Homework Data

Item	Total number of responses	Number of coded responses
1	1,156	250
2.a	531	100
2.b	541	100
3	314	100
4	1,048	100
5.a	533	50
5.b	532	50
6	987	100
7.	990	100
8	1,004	100

Note. 2.a and 5.a are the versions of the question without a context.

Since each question in the study (except the first question) was formulated so that one or more informal notions would not be useful (see Table 2), analysis of student responses had the potential to provide information on students' alternative ways of reasoning about variability. For instance, the second question in Appendix B was designed to elicit students' alternative ways of reasoning when the datasets to be compared have equal ranges. The three datasets have equal ranges, so a common student response could be that these three datasets have equal variability. Students may also suggest some informal notions, such as the last dataset is more variable because the data values are "evenly" distributed. Alternatively, students may correctly identify the set with the most clustering around the center.

In addition to these informal notions, I also focused on the role of the presence of a context in this analysis. I investigated whether or not there was a difference in the sophistication of students' reasoning in tasks with a context versus those without a context. Thus, half of the students in the course (chosen randomly) took some of the questions (see Table 2) with the context, and the remaining half took them without a context. I reported the results of the analysis in terms of variety and frequency of common ways students reasoned variability in each question, and an accompanied summary of the general trend in students' answers.

Pilot Study

For piloting the interview part of the study, I contacted one of the teaching assistants in the Department of Statistics during Spring 2016 and asked her to announce my research study to the students in her introductory statistics course sections. A few days later, she sent me a list of students who were willing to participate in my research. I contacted those students and arranged times for the interviews. I was able to pilot my first interview protocol with three students in the week of March 28, 2016 and the second interview protocol with two students in the week of April 4, 2016.

The interview protocols of the pilot study included background questionnaire and open-ended questions on variability, and different statistical tasks for each interview. The interviews started with a brief conversation about the interview participant's educational background, followed by the participant's descriptions of variability, and eventually focused on the statistical tasks. The interviews were both audio and video recorded. Four of the interviews lasted 60 to 70 minutes; only one student needed less than 50 minutes in order to complete all the tasks. The pilot study suggested that the interview tasks could generate valuable data for me to seek answers for the research questions of the study.

The next task after conducting interviews was viewing the video recordings in order to examine how each participant reasoned about variability. Based on this experience, I revised some of the tasks and follow-up questions. For example, I generated additional distributions for tasks, calculated summary statistics for the distributions to be given to the participants upon their request, added more follow-up questions, and reordered the tasks. Overall, the pilot study provided insights about the data collection and the appropriateness of interview tasks and follow-up questions for the study. For instance, I found that the interview participants became less reluctant to discuss their statistical learning in their classes after they observed that I was not evaluating their performance. Because of the same observation, they also seemed to be more relaxed and comfortable in answering the follow-up questions and noticing that they held some contradictory notions about variability.

CHAPTER 4

RESULTS

In the previous chapter, I introduced the research setting, participants of the study, data resources, data collection process, and finally, data analysis. In this chapter, I present the results of the study. First, I provide an overview of the results. The results specific to each research question are described after the overview. Finally, additional findings on students' reasoning about variability for categorical data and bar graphs are provided.

Overview of the Results

This section begins with an overview of the results of the analysis of students' responses to homework questions. Next, I summarize the findings from the analysis of the interview video recordings.

The analysis of homework data suggested that students often included information about the shape of the distributions in their responses, but the responses fell short in addressing variability specifically. I present the results of the first homework question to exemplify the overall approach that students in STAT 2000 demonstrated in their responses to homework questions.

In the first homework question, students were asked to describe a histogram that showed the distribution of the number of ounces of coffee that had been drunk by each of a random sample of 237 college students. The goal in asking the question was to investigate the extent to which students addressed variability in a typical "describe the distribution in a histogram" question. Students were expected to mention the shape, center, and variability of the distribution in the context in which the data were given. According to the histogram, the majority of the observations were clustered between zero and 30 ounces of coffee. The distribution was right-skewed and bimodal. Two distinct groups were evident in the distribution: one large group of observations in which most of the students consumed coffee less than 30 ounces, and a second smaller group of observations in which the amount of coffee consumed by students were clustered between 60 and 90 ounces. Variability of the distribution could be addressed specifically by including explanations based on the range of the distribution and determining whether the observations, on average, took similar values or excessively deviated from the average coffee consumption.

My coding of a randomly selected sample of 100 student responses according to Arnold's (2013) distribution framework suggested that students rarely mentioned variability in their descriptions. The student responses often included information only about the shape of the distribution: 62 out of 100 students mentioned that the distribution was right-skewed. Another common observation among responses was that the distribution was bimodal (15 out of 100 responses). Although the terms *skew* and *bimodal* could suggest some indirect information about the variability of the distribution, I found the use of these terms by the students to be more about the shape of the distribution than about its variability. It might, however, be the case that students had assumed that providing explanations about the shape of the distribution also implies information about the variability of the distribution. For example, students might believe that if a distribution is skewed, then it also is more variable. From the perspective of such students, having only the information that the distribution was right-skewed addresses both the shape and the variability of the distribution.

Observing that only a few students addressed variability explicitly in their responses, I continued to code responses. As a result, I ended up coding 250 student responses following Arnold’s (2013) distribution framework. The results suggested that 75 (30%) student responses provided pieces of information that could potentially be categorized as responses about variability. These responses mostly fell into either “interval for high/low for groups” or “majority” features of a distribution according to Arnold’s framework (see Table 3 for the full list of features). In addition, the coding suggested that a distinction between these two categories was difficult and unproductive to make since many responses could be coded either way. As a result, I collapsed these categories into a single category and named the category “addressing variability.”

Because the main purpose of analyzing the homework data was to explore how students addressed variability in their responses, and since most of the students did not include specific variability terms in their responses, I searched for some key terms in all 1,156 responses. Next, I accumulated a list of terms based on the student responses that I had already examined. These terms were (a) “range,” (b) “spread” (so that both spread and spread out would be captured), (c) “var” (in order to include the responses that used varies, variation, variability, (d) “deviat” (in order to include the responses that used deviate, deviation, or standard deviation), and (e) “cluster.” Table 5 summarizes the frequencies and an example for each of the terms. As Table 5 shows, only a small portion of students (47 out of 1,156 responses) used a more normative language to address variability. Within these answers, the use of *range* and *spread* were the most common.

Table 5

Summary of the Use of Variability Terms by 1,156 Students

Term	Count	Example
Range	20	Looking at the histogram, the distribution is skewed right. The mode of the data is 0, while the median of the data is around 50. To describe the spread, the range is 100.
Spread	12	Shape: the histogram is sort of skewed right but looks bi-modal with a less pronounced second mode. Spread: in the case of "Ounces of Coffee Per Day", the values definitely seem significantly spread out. Outliers: doesn't necessarily have any pronounced outliers.
Var	7	The distribution of the number of ounces of coffee college students drink is somewhat varied. The majority of college students, 100, drink around 0-10 ounces of coffee a day.
Deviat	6	This graph has a wide distribution and has a large standard deviation. It has an outlier of an unusually small observation. Most college students do not drink more than 20 ounces of coffee per day.
Cluster	2	The majority of students drink between 0 and 10 ounces of coffee per day. More students seem to drink less coffee compared to more as indicated by the cluster of students that drink between 0 to 30 ounces per day. It then drops off after 30 and only rises slightly again from 60 to 90 ounces.

Because only a small set of the student responses addressed variability through the use of statistical terms, such as *range* and *standard deviation*, I reexamined the responses in hopes of finding other possible ways of addressing variability that I might have overlooked. Another distinct group of responses that was common but did not fit into the categories in Table 5 or the features in Arnold's distribution framework was the responses that summarize the frequencies for each bin in the histogram. The following student response is an example of this type of response:

> Of the sample, 100 students drink 0 ounces of coffee, 50 drink 10 ounces of coffee, approximately 27 students drink 20 ounces, less than 10 students drink 30 ounces, less

> than 5 students drink 40 ounces, less than 5 students drink 50 ounces of coffee, approximately 10 students drink 60 ounces of coffee, around 12 students drink 70 ounces of coffee, less than 15 students drink 80 or 90 ounces, and less than 5 students drink 100 ounces.

Although one might claim that these types of responses address variability, I did not regard them in that way. A histogram emphasizes an aggregate of data; thus, the description of a histogram should capture the general features. The example above shows a listing of individual facts rather than an overall summary. In sum, variability is described as the characteristic of a distribution (Ciancetta, 2007), which was missing in the students' explanations.

Overall, the analysis of student responses to the first question suggested that for more than half of the many students, skewness was the most recognizable attribute of the distribution, and students often mentioned skewness but nothing else in their responses. In addition, available student explanations that included variability were often too brief to produce fully elucidated insights about students' reasoning about variability. Many students often used terms, such as *spread,* in their responses but did not explain in detail the way a distribution was spread. For example, a student's response for the second homework question was as follows: "The definition of variability is the spread of data, and it's spread out equally." As the response indicates, the student might have assumed that using the term *spread* is self-explanatory when addressing the concept of *variability*. Accordingly, the results of the analysis of student responses was sometimes less informative than I had hoped in seeking answers to the research questions of the study. The rest of this section presents the interviewed students' reasoning about variability.

As I explained in the last chapter, the interviews started with a brief dialogue about how the student would describe variability in general. Words and expressions that the interview participants commonly included in their responses were "change, different" (Ocean), "all possible choices, nuance, differences" (Karen), "vary from each other" (Chloe), and "different ranges, different numbers, different results" (Britney). Overall, the participants' descriptions were more applicable to the colloquial use of the word *variability* than to its formal use. The participants did not confine their use of *variability* to the statistical meaning. I concluded that the descriptions that the participants devised at this phase of the interviews were reasonable because the aim in these brief dialogues was to capture the participants' broad understanding of the *variability* concept not the meaning of *variability* that is confined solely to the discipline of statistics.

Later in the interviews, two of the participants expanded their definitions of *variability* to include the notion that the statistical concept of *variability* might have a different meaning if compared to its use in daily life. Ocean pointed out, "*In math situations* [emphasis added], variability has a totally different definition." Similarly, Chloe claimed that the term *variability* could have distinct meanings in statistical and non-statistical situations. For example, she explained the distinction between the meanings as follows:

> The students of this campus vary within each other and also from other college campuses. There is always different, sort of, different student population that's what I think of variability as a whole. Now, of course, *I am in statistics*, I know that it means how one sample varies from the other. It's basically showing how they are not the same but they could be similar in a way…yeah…there is always differences between each person and each sample.

For the various statistical tasks that followed these brief conversations, Ocean stood out as the only student who consistently focused on where the majority of the data were in relation to the mean of the data. She addressed my follow-up questions using this idea throughout the tasks in the interviews. In addition, she successfully considered range, extreme values, mean, and other possibly related issues when reasoning about variability. She was not concerned that certain tasks were designed to focus on specific aspects of the distributions (thus, the tasks might trigger some of the aforementioned informal notions). In addition, if she had contradictory conclusions in comparing variability across distributions, she handled those situations based on her fundamental ways of reasoning about variability instead of adopting a task-specific approach. In other words, Ocean's powerful conceptualization of variability as the measure of clustering around the center allowed her to negotiate possible conflicts that her counterparts were unable to navigate during the interviews. The following excerpt illustrates the general mechanism that she employed when she addressed variability:

Ocean: Initially I was thinking of variability as…just like…the *different data points* [emphasis added] so…like…something is more variable if different data points were recorded with all different but here you can see that they were, like, three of them were all the same but still constituted for the higher variability…it is not always have to be. Each individual data point does not have to be different; it just *depends on how far each varies from the mean* [emphasis added].

Oguz: So, then would you reconsider your very first list of words you gave to explain variability, change and different?

Ocean: Different. Yeah, I wouldn't use that anymore, because as I said here they are all the same, but they still have a higher standard deviation. I don't know… I guess...like…in *a math situation* [emphasis added], *variability* had a totally different definition. Using it, like, just using it like a common language like talking to your friends, "that varies"; that's what I like is changing, but here it really doesn't have to change but still had variability within the dataset.

Moreover, context was not a mediating factor in Ocean's reasoning about quantitative variability. She did not alter her approach to variability across the context-included and context-free tasks. The only major difficulty she encountered was on Task 5 of the second interview. In Task 5, Ocean was asked to reason about variability for categorical data given in bar graphs. The first question of the task included two bar graphs that showed the frequency distribution of four categories in two different samples. Ocean was asked to choose the more variable sample and explain the reason for her choice. The second question of the task was a similar but simplified version of the first question, and the third question was another bar graph representation of categorical data for four grade levels' preferences about where to go for field trip: the aquarium or zoo. Overall, Ocean had difficulty in reasoning about variability in these situations. Her strong reliance on the idea of variability as *clustering around the center* seemed to limit her reasoning with categorical data as the following excerpt suggests:

> My explanation would be like with the mean and stuff, but you can't calculate a mean from these [points to the categories in the bar graphs]. Because that was when I first think of variability. Now, I think of like average and stuff like that and changed the first definition that I gave you.

Although other interview participants occasionally attended to the notion behind standard deviation (especially Karen later in her interviews), their ways of reasoning were less consistent across the tasks of the interviews. The participants employed the notions *range* (e.g., in Chloe's words, "range has a lot to do with variability") and *differences of data values* (e.g., in Britney's words, "variability means the most different") as justifications for their decisions on variability. They usually attempted to use these two criteria whenever possible as the following excerpt from Chloe's responses to Task 4 of the first interview illustrates:

Chloe: Maybe it doesn't have to do with range sometimes. This is ten and fifty and that's forty [points to the first dataset]. And it's a difference of forty, and this is a difference of sixteen [points to the second dataset]. However, this is repeated three tens, three twenties, and three fifties, that's not very variable. You know it's like ten, ten, ten, then like twenty, twenty, twenty, then fifty, fifty, fifty. That's not much variability at all.

Oguz: Is this what you have in mind?

Chloe: Yeah, I would say that second one has more variability. Maybe range does not have in terms of variability because this range is a lot bigger, but if you got repeat so…like…I don't know, I just…I don't know how to describe that change. Now you have one person who thinks this, who thinks that [inaudible], you have more varied responses. These are the very similar responses [refers to the first dataset], whereas this one has very similar responses—three people here. Yeah, sometimes range doesn't matter, frequency matters.

Oguz: Frequency matters.

Chloe: *Frequency and range matters* [emphasis added], and in this set both matter, but frequency I think overpowers because you are getting a lot of the same here and there and there.

Oguz: Okay.

Chloe: Doesn't give you much of a difference. It's the same amount of people, too. Three people here and there. That's not too different but in there [inaudible] two people here, one person would be here. It gives you a broader range of data to work with.

Those three interview participants also directed their attention toward the clustering of observations in a given distribution, especially when their informal notions were less applicable to the given situations. It should be noted, however, that the treatment of clustering as used by the participants was still naïve because they generally failed to consider the position of clusters of data in a distribution in relation to its center. After I observed that the interview participants failed to include the mean in their reasoning about variability I provided graphs in which means were labeled and asked if they could describe the relationship between the mean and the variability of the distribution. Britney, for instance, confirmed that she was not able to connect the information about the mean of a distribution with the distribution's variability. She said, "I am not sure about the relationship with, how points close to the mean has any relationship to variability." These results, together with the results of the second homework question, presented in the next section, suggested evidence of the difficulty students tend to have with incorporating the location of the center into their thinking about variability.

In the third question of Task 4 in the first interview, two raw datasets were given, and subjects were asked to determine whether the first or second dataset was more variable or if they were both approximately equally variable. This question was among one of the few instances in which interviewees were not able to use either range or the different data values ideas. In this case, two of the interview participants decided that the distributions were equally variable. Those participants determined that the two distributions had the same range and same variety of data values so (they concluded that) the distributions were equally variable.

It was noteworthy that the interview participants often failed to reason consistently across different tasks and questions. They switched between different ways of reasoning based on the affordances and restrictions that were caused by the particular characteristics of the tasks they were asked to perform. They employed informal notions in different weights in different questions, which was probably because the characteristics of the distributions or datasets provided in the tasks underlined certain aspects over others. For instance, in one of the tasks, Karen commented, “In some cases, when the standard deviation is high, there is more variability, but in this instance I don’t think it applies.” Similarly, for Task 5 of the first interview protocol, Britney had two contrasting ways of reasoning, but she reported that she was unsure how to decide between them:

> Honestly, I go back and forth on two [the uniform one] and three [the u-shaped one] because I could kind of see either way. Because I did this, because look how it is evenly spread out, so there is a lot of variability … each of them own the same frequency interval. But in this one, it has a big chance of ten and zero ranges, kind of.

Overall, the inconsistencies in interview participants’ approaches to the concept of *variability* did not seem to bother them. The rest of the section provides answers to each research question of the study based on the analyses of homework and interview data.

Results Specific to Each Research Question

Research Question 1.a:

How do undergraduate students reason about variability when the datasets or distributions to be compared have equal ranges?

Most of the homework questions could be exploited to answer this research question because the distributions in those questions had equal range values. The results for two of these questions, the second and the third, are presented in this section.

In the second homework question (see Figure 1), three distributions X, Y, and Z, each contained five values, were given and students were asked to choose the least variable dataset from those distributions. Although the distributions had equal range values, the observations in Y were more closely clustered around the mean of the distribution, but the observations in X were clustered away from the mean. Hence, the distributions from the least to the most variable were Y, Z, and X, if students regarded variability as the measure of how much, on average, observations in a distribution deviate from its mean.

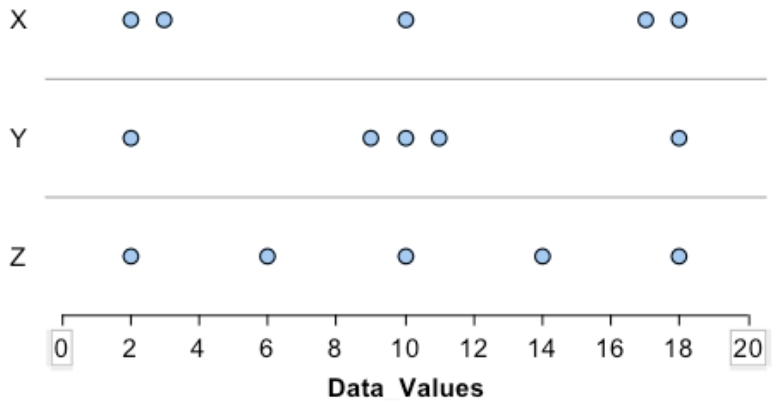
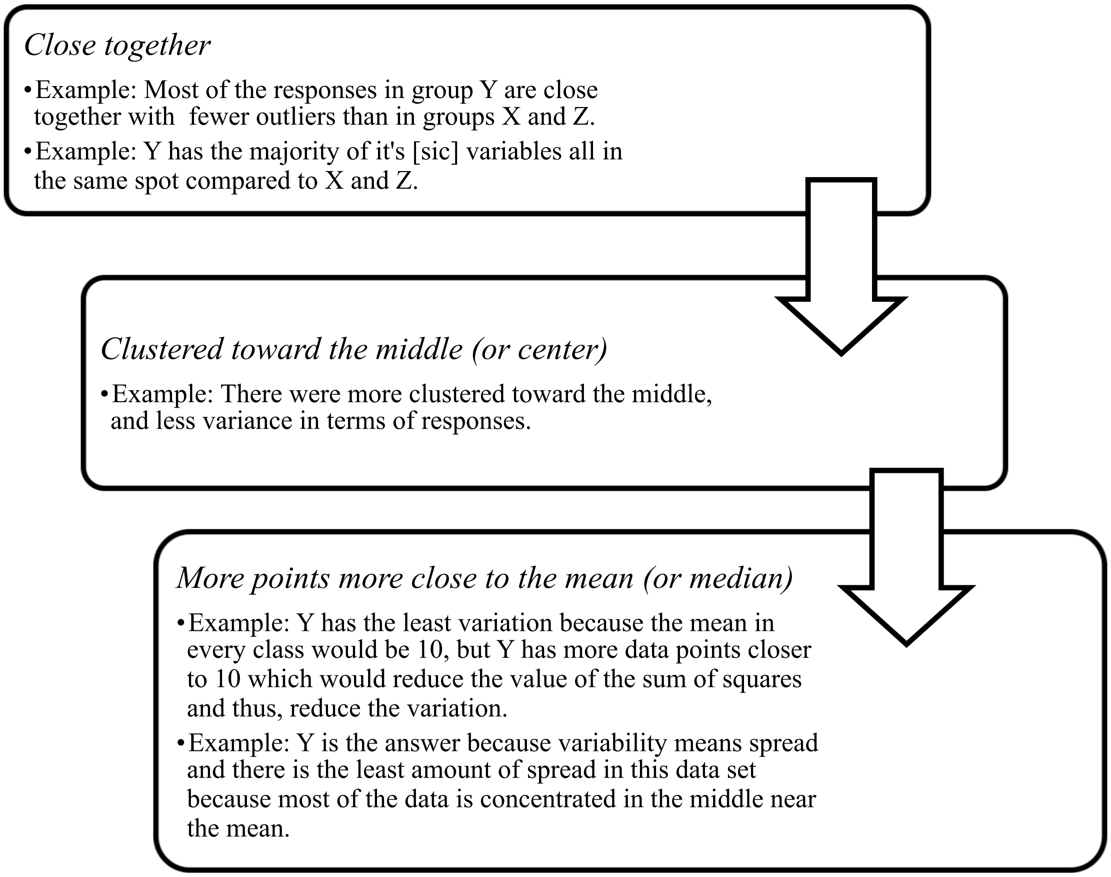


Figure 1. Homework question 2.

The majority of the students chose either Y (41.6% of the responses) or Z (38.7% of the responses) as the least variable distribution. Although choosing distribution Y as the least variable was the expected, normative response because the observations in Y were more closely clustered around its center, 415 out of the total 1,072 students chose Z as the least variable distribution. As a result, for the rest of the analysis, I focused on the justifications students provided for choosing either Y or Z.

Students' justifications for selecting Y as the least variable distribution suggested that the main reason why students chose this option was the *clustering of observations*. In order to explain their reasoning, students commonly used phrases such as *more condensed*, *clumped together*, *concentrated in the middle*, and *less spread out*. It should be noted that the language students used in this question was not similar to the features of variability listed in Arnold (2013), so I coded their responses independently of Arnold's framework.

An examination of the explanations for this answer choice Y suggested that the responses differed based on whether the responses explicitly included the measure of center. Therefore, I re-read the responses and distinguished the set of responses that mentioned clustering but did not take center into account from the set of those that clearly included the center in the explanations. Coding the responses according to that distinction suggested that these responses could be categorized in three distinct groups. Figure 2 displays these reasoning themes from vague to more deliberate consideration of center, with examples for each theme.



Close together

- Example: Most of the responses in group Y are close together with fewer outliers than in groups X and Z.
- Example: Y has the majority of it's [sic] variables all in the same spot compared to X and Z.

Clustered toward the middle (or center)

- Example: There were more clustered toward the middle, and less variance in terms of responses.

More points more close to the mean (or median)

- Example: Y has the least variation because the mean in every class would be 10, but Y has more data points closer to 10 which would reduce the value of the sum of squares and thus, reduce the variation.
- Example: Y is the answer because variability means spread and there is the least amount of spread in this data set because most of the data is concentrated in the middle near the mean.

Figure 2. Student explanations from less clear to clearer consideration of center.

To claim a distribution as less variable, the observations in the distribution need to group closely around the center of the distribution. Accordingly, students' use of terms such as *middle* and *center* made it difficult to decide whether these terms were used to refer to the midpoint of the values on the *x*-axis or the mean or median of the distribution. Nevertheless, these responses suggested a clearer emphasis on center when compared with the group of responses in the top group in Figure 2. In conclusion, although the notion that "there are more values closer together" is important to observe, a more complete response should include more clarification as the following student response suggested: "Section Y has most of it's [*sic*] data clustered around the center of the dataset, leading to less variation in the data."

When students chose dataset Z as the least variable dataset, they commonly used expressions such as *equally spaced, evenly distributed (or spread), uniformly distributed, consistent (constant, even),* and *spread*. An overwhelming majority of those student responses included an argument that Z was the least variable because the observations in Z were "equally spaced" and thus had "greater consistency and lesser variability." Treating variability in this way seemed to be a common way of reasoning among students as the following responses illustrate:

> I chose Z because it has a continuous flow of values. Each number is 4 units apart, and the variability stays the same throughout the whole line of numbers. The other parts have much more space between the depicted numbers, so there is more variability in the line.
>
> Because the values in Z appear at constant intervals and are very organized. They always appear at intervals of two, giving room for little to no variability.
>
> All of the variables are separated by 4, so there isn't any variability among them.
>
> There is few variation between each of the point.
>
> Z is evenly spaced and is constant. There is no data variability.

Students' justifications for choosing the distribution Z suggested that students hold an informal notion of variability, triggered by the characteristic of Z. Instead of focusing on how the data values were spread from the mean value, students compared whether the overall spread was more haphazard (as in X or Y) or followed a regular pattern (as was the case in Z). In other words, they perceived (smaller) variability as consistency or predictability of the spread of values in a dataset. The following student explanations clearly demonstrate this informal notion in detail:

> I chose graph Z because although the dots are more spread out than the other two, it is a constant gap between the points. It varies the least because it begins at 2 and every fourth number after that has a point. It's systematic. X & Y may have points that are closer together, but their *spread is more random* [emphasis added] while Z is consistent.
>
> Z has the least variability because variability refers to how points differ from each other using range, standard deviation, and variance, which is all the same regardless of which point it is since the points are spaced evenly apart and have no outliers.
>
> Variability means have a lot of difference within data. Dataset Z has each value in an equal distance away from each other showing a constant standard deviation while the other datasets have unequal distributions showing variation within the data.
>
> Variability can be defined as "having a lot of differences" or "spread out". The data in dataset Z has similar distances between each dot compared to the other plots. The other plots have a much greater space between some dots.

As the responses above clearly indicate, these students focused on the extent to which consecutive data points were evenly distant from each other. In other words, the existence of a regular pattern was satisfactory evidence for students to claim that the distribution Z was the least variable. This informal notion could also be framed as the consistency or regularity of the distance between data values. Note also that students' informal notions of variability seemed to be dependent upon the properties of the distribution given to the students. This question, for example, led to reasoning that was not observed for the other homework questions in the study.

Lastly, although most of the students chose either dataset Y or Z as the least variable, 12% of the students in this study concluded that the distribution with the least variability could not be determined since the range values were equal across the given distribution. Here is a typical justification one student provided for this conclusion:

> I chose that it can't be determined due to the having the same range and variability all of the datasets *start at 2 and end at 18* [emphasis added].

The student response above and other similar ones indicated that students equated variability with range and concluded that X, Y, and Z should have equal variability. Although the literature indicated that students usually include range when reasoning about variability, having only 12% of the students in this study relying on range was noteworthy. The result suggested that students may reconsider their use of range based on the constraints and affordances of the question, dataset, or distribution when reasoning about variability.

The third homework question required comparing the variability of test scores between two classes, Class A and Class B, represented in histograms. Note that both histograms had a range of 100 points. My coding of 100 randomly selected student responses according to the features in Arnold's framework suggested that a large proportion of the students addressed the shape of the distribution specifically. These students usually suggested that the distribution for Class A was right-skewed and the distribution for Class B was closer to a symmetric or normal distribution. Table 6 provides the counts and students' typical justifications for their conclusions.

Table 6

Percentages and Examples for Each Category of Response in Question 3

Category	Count	Example
Class A is more variable	41	The range of test scores for both class A and class B are the same in this example so *we have to consider the frequency of the test scores to determine variability* [emphasis added]. Class A test scores follow a right-skewed distribution while class B test scores follow a normal distribution. Class A *frequencies vary* [emphasis added] in magnitude more drastically than class B frequencies.
Class B is more variable	16	Overall, Class A's test scores are far less spread out than Class B's test scores. This is because many of the students in Class A scored between a 0 and 30 but most of the students in Class B scored between a 30 and 90.
Class A and B equally vary	4	The variability of the test scores are [*sic*] the same because the spread of the data is the same. The ranges are the same for both classes. The difference is that Class A's data is skewed to the right while Class B's data is approximately bell shaped, maybe SLIGHTLY skewed to the left.
No discussion of variability	39	Class A's test scores are very low and the histogram is skewed right. The mean is a low grade. Class B's test scores were much higher and the histogram is skewed left. The mean is a higher grade.

As shown in Table 6, the majority of the students either chose Class A as more variable or did not mention variability in their answer. Note that the question specifically asked the students to compare variability between the given distributions, but a large proportion of sample student responses did not include a discussion of variability. These responses generally included information about the shape of the distributions using the term *skew*. One student, for example, suggested, “The variability for Class A is skewed left, while Class B is a normal curve.”

As I elaborated in the previous section when reporting the results of the first homework question, it was possible that students use the information about the shape of a distribution to address variability as well. In other words, some students might have assumed that the term *skew* was self-explanatory in addressing the variability of the distributions, as the following examples suggest:

> *The variability of the distributions of test scores* [emphasis added] for Class A are skewed to the right and the distribution for Class B is skewed to the left.
>
> Class A is skewed right where as Class B is more bell shaped with a skew to the left. *The skew is not as prominent so there is more variability* [emphasis added] in Class A's scores.

The students' justification for the decision that Class A was more variable indicated a common misconception reported in the literature for histograms—associating differences in heights of bins with larger variability. In other words, students often assert that if the "frequencies vary" in a distribution then the distribution is more variable. Similarly, as the following student response suggests, students often claimed that variability of distributions could be judged based on "how similar or different the frequencies" were in a given distribution:

> The variability in the left graph (class A) is much greater than the variability on the right (class B) because there is much greater differences in the frequencies while on graph B the frequencies change much less drastically.

The analysis of responses did not indicate a serious consideration of the center when reasoning about variability. Responses usually lacked mention of the mean or other measures of center. Ten out of the 100 analyzed student responses explicitly mentioned the fundamental idea of variability using more appropriate statistical terminology in their responses. The following student responses were examples from that group of answers:

> The scores for Class A show a curve with a skewed-left distribution while the scores for Class B show a curve with a skewed-right distribution. The mean and the median were far higher for Class B's test scores than for class A's test scores. The variability for Class A is higher because it has points that are farther away from the mean than does Class B.
>
> Class A has less variability of test scores because a majority of the data is around the class average, indicated by the high frequencies clustered together. Thus, the data has a small standard deviation. Class B has high frequencies in the various test scores, not just

> around the mean, indicating a larger standard deviation. Therefore, Class B is more variable.
>
> There are 4 different things we need to consider for variability; range, mean variance and standard deviation. First the range is the same for both of them, but as for the mean Class A is closer to the 20 and for Class B it should be around the 50s. For standard deviation, the Class A score will have a larger range compared to Class B.
>
> There is less variability in the test scores for Class A than there is in the test scores for Class B. This is because Class A's test scores are less spread out; the scores are more concentrated toward the lower end of the grading scale. Class A’s distribution is skewed right. There is less deviation in Class A’s test scores. Class B’s distribution represents a relatively normal distribution. Class B’s test scores have more variability because the scores are more spread out, with no strong grade concentration and greater deviation.

Overall, the analysis of the second and third homework questions suggested two informal notions of variability. Many students interpreted variability as the extent to which consecutive data points were evenly distant from each other in the second question, and many students interpreted variability as the extent to which frequencies were different in the third question.

The analysis of interviews suggested mixed findings for datasets with equal ranges. Of the four interview participants, Ocean was the only student who focused her reasoning about variability on clustering around the center and standard deviation. Although she initially attempted to coordinate the notions of “larger range means larger variability” and “clustering around the center implies less variability,” she dropped the former and relied on the latter more uniformly throughout the remainder of the interviews. The other participants demonstrated various reasoning approaches discussed below.

In Task 2 of the first interview protocol, which was identical to the second homework question except the task had multiple follow-up questions, Chloe, Britney, and Karen suggested that they initially assumed that range was the most prominent measure when reasoning about variability. They were willing to adopt a different approach, however, when they observed that range did not provide distinguishing information about variability because the compared groups had the same range. Although they articulated their thinking in slightly different ways, these interview participants commonly suggested an informal notion: The distribution with more distinct values was the most variable. As a result, if they observed any clustering (irrespective of the relative position of the cluster to the center) in a distribution, they claimed that the distribution was less variable.

The interview participants articulated the informal notion in slightly different ways. For example, in order for Chloe and Britney to judge a distribution to be variable, the observations needed to differ from each other, which also required the observations to be far apart. Accordingly, clustering indicated less variability because clustering signaled that the observations were very similar (in terms of values) to each other. Karen put this informal notion into words as “covering more data values,” and as long as the observations of a distribution were different from each other, she noted more variability. Therefore, Karen noted that if the observations were “close together, too, they are not gonna cover the most ground;” consequently, the distribution would not vary much.

This informal notion was more prominent when the interviewees were asked to work with the datasets that were given in raw data form. When the interviewees observed that the datasets to be compared had equal ranges, they raised the issue of whether the datasets contained repeating data values. For example, all interview participants except Ocean focused on how different the data values were in the first question of Task 4 (in the first interview protocol). In this question of the task, two raw datasets with equal sample size were given, and participants were asked to choose the dataset with more variability. The participants decided that the second dataset was more variable since all nine observations in the second dataset were different numbers as opposed to the first dataset, which had three different repeating numbers. The fact that the first dataset had a relatively larger range did not prompt the interviewees to conclude that the first dataset was more variable. Those participants again chose the second dataset to be more variable in the second question of the task (in which both datasets had the same range value) because the second dataset had more varied data values. Overall, the interview participants' approach to these and other similar questions in the interview tasks suggested that the use of a "variety of numbers" was an influential way of thinking about variability when they were unable to use range in their conclusions.

The interview participants' use of this informal notion took a different form when they worked on comparing variability using conventional statistical graphs, such as dot plots and histograms. Although Ocean maintained her original approach—gauging clustering around the mean—in these tasks, the others suggested different criteria in order to claim that a dot plot or histogram was less variable. Britney and Karen claimed that the approximately uniform distributions would be more variable when two histograms or dot plots had equal range. When Chloe worked on the Task 5 of the first interview, she decided that uniform distributions were less variable because she claimed that the presence of different frequencies in dot plots or histograms caused larger variability. Accordingly, she claimed that normal distributions were more variable than uniform distributions when both had the same range values.

Overall, the analysis of second and third homework questions suggested two informal notions of variability. In the second question, many students interpreted variability as the extent to which consecutive data points were evenly distant from each other, and in the third question, many students interpreted variability as the extent to which frequencies were different. Considering the interview results, Ocean was the only interview participant who used the notion of clustering around the mean when she compared distributions in terms of variability in the case of equal ranges. The other participants usually used presence of various data values in raw datasets in order to identify a distribution as more variable. The use of these strategies by the interview participants suggested how this informal notion was prominent in their conceptualization of variability.

Research Question 1.b:

How do undergraduate students reason about variability when the datasets or distributions to be compared have no extreme values?

The research question was about investigating students' reasoning about variability when the datasets or distributions to be compared have no extreme values. The third homework question, whose results were reported in order to address Research Question 1.a, could also be exploited for Research Question 1.b. As discussed above, the distributions for Class A and Class B did not have extreme values, or at least, the students did not claim there were any extreme values in the given histograms. Accordingly, no students presented a justification based on individual or extreme values. As discussed above, the students focused on the frequency differences within the bins of a histogram and compared it to the case in the other distribution.

Homework Question 4 also targeted research question 1.b. because the first part of the question had two extreme values (observations 8 and 10), whereas the second part of the question lacked those extreme values. Students were asked to describe the distribution in both cases so that the role of extreme values in students' reasoning could be investigated. My coding of 100 randomly selected student responses revealed that 34 of those responses included the terms *extreme values* or *outliers*. Among those student responses, half of the students used the existence of outliers to support their ideas about the shape of the distribution (e.g., the distribution is skewed since 8 and 10 are outliers). In other words, those students used the presence of the outliers (or extreme values) to justify skewness but did not mention the outliers' role in variability. The other half of the students used outliers to justify their conclusion about the variability of the distribution. Examples for each purpose are presented below:

> It is relatively symmetrical with the exception of a few outliers, which cause a slight skewed left overall [use of outliers to discuss the shape of the distribution].
>
> Thew [sic] range is now 4 instead of 10. The mode remains two pets, but the chart reflects more of a bell curve shaped distribution. The range and variability decreased [use of outliers to discuss the variability of the distribution].

No students claimed that the distribution was variable based on the presence of outliers, leading me to conclude that they did not consider outliers and extreme values when they reasoned about variability. Thus, I concluded that availability or unavailability of outliers was not extensively influential in students' reasoning about variability.

My coding of 100 randomly selected student responses according to the features in Arnold's framework also suggested that some students partially addressed the important characteristics of the given distribution. An exemplary student response, which was rare among the coded responses, is as follows: "The distribution of the number of pets owned by students is mostly centered around 2, and is slightly skewed to the left, but with some outliers to the right of the center of the data." As the response clearly shows, the student estimated the center and shape of the distribution, mentioned where the majority of data were in relation to the mean, and identified possible outliers in the given context.

Out of 100 responses, only 35 addressed variability in their descriptions. Among the set of responses that addressed variability, students often used expressions that fall into Arnold's *Majority (mostly, many)* feature. These responses often provided descriptions by using the phrases *most* and *majority* as the following examples illustrate:

> Most people have about 2 pets, usually between none and 4. Very few have more than that.
>
> The majority of class members owned between 0 and 4 pets. The students who owned 8 and 10 pets can be counted as outliers.

Many responses (65 out of 100) were not informative enough to conclude whether students had considered variability in their responses. For example, in responses such as, “The distribution is skewed right. This means that more people own fewer pets, like 1 or 2, rather than a lot of pets, like 10” or, “The new distribution would be skewed left because most of the data would be on the left side of the peak, which would be 2,” the students’ main ways of reasoning were not clear from their explanations. In other words, student responses were too vague and ambiguous, which made it too difficult to understand how they reasoned about variability in the given homework question.

The sixth homework question included two dot plots that showed the number of pairs of shoes owned by females and males who took a survey. My coding of 100 randomly selected responses suggested that 66% of the students concluded that the distribution for females was more variable and 26% of students concluded that the distribution for males was more variable. Students generally used two phrases *more spread out* and *more clustered* in their responses for either of the conclusions. For example, students who concluded that the distribution for females was more variable used the phrases *spread out* and *more clustered* as the following example illustrates: “The samples for Females are more variable because the samples are more *spread out* [emphasis added]. The samples for Males are more clustered together within the range and the Females samples are all very dispersed within the range.” Similarly, one student provided the following explanation for his or her conclusion that the distribution for males was more variable:

"There is more variability in the male group because the dots are more *spread out* [emphasis added]."

When students used terms *spread out* and *more clustered* without additional explanation, I could not classify the student responses further using the features in Arnold (2013) because the responses were not detailed enough to understand what students meant by these terms. For example, one student claimed, "There is more spread for the females, therefore, more variability," which left me unable to delineate the student's reasoning.

Analyzing student responses also suggested use of some other common phrases. One of the phrases that students used was *concentration*. Students' use of *concentration* in their explanations might indicate that they treated variability as deviation from the mean. It should be noted, however, that use of the word *concentration* does not assure they had such an idea. Another common occurrence was the use of *skew*. Some students also used the term as if skewness was a direct and clear measure of variability. Accordingly, student explanations such as "the distribution for females is more variable because the responses are more spread out whereas for males the responses are skewed right" were not clear enough for me to decide how these responses indicated variability. In contrast, the following example provided a clear link between variability and skewness: "I said that the distribution for Males is more variable, because it is more skewed and would have more points farther from the mean." As evident from the last part of the response, the student explained how skewed distribution could result in increased variability.

Other than describing the distribution with the phrases *more clustered* and *spread out*, only a few students explained their reasoning using the terms *outliers*, *standard deviation*, and *range*, and explicitly addressed the center of the distributions and where the majority of data were in relation to the mean. Thus, the following types of explanations were rare among the set of responses I analyzed:

> I said that the distribution for Males is more variable, because it is more skewed and would have *more points farther from the mean* [emphasis added].
>
> The males have more variability because while both sets of outliers that skew them to the right, there is *more data clustered near the median* [emphasis added] of the males [*sic*] distribution than the females.

Overall, the analysis of the sixth homework question suggested that students generally used phrases such as *spread out* and *more clustered* without closely considering whether using them accurately conveyed their ideas about variability. In addition, the result suggested availability and unavailability of outliers was not influential in students' reasoning about variability.

The analysis of interview data also did not suggest that availability or lack of availability of extreme values had a necessary role in the interviewees' reasoning, although they occasionally raised concerns about the probable impacts of outliers on variability in given distributions. Ocean's work in Task 3 of the first interview (the task with a dot plot that represented the number of pets owned by each of 30 students) suggested that she did not neglect possible extreme values, but she also did not overemphasize their role in terms of reasoning about variability. Ocean's first observable reactions to the task were that the data values were "clustered near the mean," and the distribution included two data values that were distinctively farther away from the rest of the data. Her response to the variability of the distribution included these two values as she claimed "Except these two, but if you don't take these into consideration and just mark off as one and two, then I would say that there is not as much variability within this distribution."

Britney, Chloe, and Karen treated extreme values and outliers in a number of ways when they reasoned about variability. Britney claimed that without outliers, “[The distribution] would be less variable because it would have a less range that it fall in between, because you would shorten the range if you cut that out.” Chloe’s reaction to the existence of outliers shared similar features to Britney’s responses. When I asked (while she was working on Task 3 of the first interview) why the presence of outliers results in a more variable distribution, she said, “Outliers increase range of values, and as I was saying in my own words, hitting more numbers.” Chloe claimed that the variability would be greater if there were outliers in a distribution because “outliers increase range of values.” Chloe’s treatment of extreme values seemed to function in this order: With extreme values, an increase in range follows, so the possibility of more different numbers in a distribution increases. She also added that variability increases more radically, especially if extreme values occur “on the side that has barely anything” in a given distribution. This observation also suggested that outliers imply more variety of numbers (i.e., data values) in a distribution because they are different values than the rest of the data.

Overall, the interview participants’ explanations suggested that they took outliers and extreme values into consideration when reasoning about variability, but perhaps as a secondary phenomenon to justify their reasoning. As Britney and Chloe’s explanations above illustrate, the primary use of extreme values and outliers by those participants was to support their reasoning based on range and variety of data values. This finding also meant that the lack of extreme values in given distributions was not sufficiently instrumental for the participants to consider variability of quantitative data in terms of how close the data were distributed around the mean. To conclude, the available interview data did not suggest a considerable difference in approaching variability between the situations in which distributions did not have any extreme values and the situations in which distributions had extreme values. In other words, the participants often preserved their preexisting notions and approaches with and without the presence of extreme values.

Research Question 1.c:

How do undergraduate students reason about variability when the datasets or distributions to be compared have approximately the same number of different values?

Analyzing 100 randomly selected student responses from the fifth homework question (comparing the variability of two datasets given in the form of raw data) suggested that the overwhelming majority of the students (84 out of 100) decided that the second dataset had more variability. The result clearly showed that most of the students were inclined to treat variability as *various data values* if one of the datasets had more varied data values than the other.

Students' justifications for concluding that the second dataset was more variable usually included an argument centered on how different the data values were between each other in the second dataset. Except for a few responses, students usually used one of two notions to justify their reasoning: *more different data values* or *spread out*. The following responses illustrate students' typical explanations for the question:

> The second dataset has more variability than the first because the first dataset only has three sets of values, while the second has multiple different values in the dataset. The ranges for both of these datasets are identical, but the second [dataset] has more variability because of the more diverse set of values.
>
> The second dataset has more variability because the numbers are more spread apart whereas the first dataset has a lot of repeats.

Overall, analyzing student responses to the fifth homework question provided limited information on Research Question 1.c. In other words, the instruments fell short in investigating students' reasoning about variability when the datasets or distributions to be compared had approximately the same number of different values.

The treatment of variability as "how different the values are from each other" was the overarching theme when each interview participants described variability at the beginning of the interviews. For example, when I asked Ocean how she could describe the term *variability*, she explained that variability refers to the mere fact that there are "a lot of different points, different data points within that population, everything is not exactly the same." She further claimed that variability basically refers to "change" or "different from normal." The words that she claimed to be related to variability were *change*, *different*, and *quantitative value*.

The notions above, however, seemed inapplicable to the quantitative data as Ocean started to work on the interview tasks. Her approach to Task 4 of the first interview (hence, to the aforementioned informal notion) was as follows:

> I expected group two [refers to second dataset] to have a lower standard deviation because when I labeled the means of those, you can see that the data points are more clustered towards mean versus this mean [refers to the relative position of the mean in the first dataset] does not have really any points near it until you go, like, further below the mean or above, extremely above the mean, so that's why standard deviation is eighteen and five point five on this one and then again smaller standard deviation corresponds with less variability where high standard deviation corresponds with more variability within that dataset.

When I raised the issue that the variety of data values was lower in the first dataset, she agreed that she "was thinking variability as different data points," but she did not use *change* anymore to describe variability since it was an inappropriate way to describe variability. She further clarified that "in *math situations* [emphasis added] variability has a totally different definition." The following excerpt illustrates how she reframed her thinking.

Ocean: Initially I was thinking of variability as just like the different data points so like something is more variable if different data points were recorded with all different, but here you can see that they were like three of them were all the same but still constituted for the higher variability it is not always have to be. Each individual data point does not have to be different; it just depends on how far each varies from the mean.

Oguz: So, then would you reconsider your very first list of words you gave to explain variability, change, and different?

Ocean: Different. Yeah I wouldn't use that any more because as I said here they are all the same, but they still have a higher standard deviation. I don't know. I guess like in a math situation variability had a totally different definition. Using it, like, just using it like a common language like talking to your friends, that varies; that's what I like is changing, but here it really doesn't have to change but still had variability within the dataset.

Oguz: Maybe there are some non-math situations.

Ocean: Yeah situation can affect the definition of variability.

Oguz: Can you think of any situation in which the notions of different or change may be an appropriate way to describe variability?

Ocean: Maybe, like I said the temperature, or like maybe amount of water bottles you drink a day varies, they are different everyday but like in math when you are thinking of variability you are not thinking of data unique by itself; it can be like repeated; it does not have to be a one-time thing.

Oguz: What about that t-shirt color example?

Ocean: Yeah like in a classroom, variability, I feel like variability can still be in the t-shirt example; it can still be used since t-shirt colors can be different in the different classrooms and then you would assume that all that to be red right after a game or something like that.

Overall, Task 4, in which the datasets to be compared in terms of variability were given in the form of raw data, specifically investigated whether interview participants might treat variability as the measure of how different the numerical values were in a dataset. Ocean realized that describing variability based on the notion of *difference* was inadequate to and inconclusive in her investigation of variability in various tasks throughout the interviews.

In terms of reasoning about variability of the distributions in raw datasets, all of the interview participants except Ocean maintained the same informal notion—possibly even more strongly as I explained above when reporting the findings for the first research question. The term *variability* essentially meant *different* and *change* for the interviewees, and the observation of various numbers in datasets seemed to confirm interview participants' treatment of variability in this way. When the notion could not be employed because of the design of the question or other features of the tasks that overshadowed this property, the participants employed different strategies that I found fragmented and inconsistent. In some of these cases, the participants classified the distributions as being equally variable. For example, upon confronting the third question of Task 4 of the first interview, in which both datasets had the same range and variety of values, Karen and Chloe decided that the datasets were equally variable. The following excerpt illustrates how Karen dealt with the situation, and eventually decided that the distributions had equal variability:

Karen: They are looking kind of similar.
Oguz: In want ways?
Karen: Well, both have repeating values in them. This one [the first dataset] has forty, forty-two. This one [the second dataset] has seventy and thirteen repeating itself so yeah and also, this I think has a similar pattern maybe because this one [the first dataset] goes up by one plus one is four, ten [in the second dataset] plus one eleven and then except for this, maybe wait, yeah maybe it's not the same yeah I don't know. This one [the second dataset] jumps a little bit more that throws variability. I think they are pretty similar. This one [the second dataset] has the highest value but and this is a big jump right here that is a really big jump. I think for the most part, they are similar.
Oguz: So they may have the same variability?
Karen :Yeah, I think they have the same variability because they are not drastically different and they both have similar characteristics.
Oguz: If we had second forty-two to be forty-three, would that make a change in your answer?
Karen: Yes, I think so because then there would be a different number right there. It wouldn't have the same amount of numbers, the mode I guess, so this one [the first dataset] would have more variability if that was the case.

For the same situation, Britney started to use the notion of clustering of values in order to justify her decision. The following excerpt illustrates how she started to establish her approach of "clustering" of values and on whether or not the presence of values in a distribution "grouped together" implied less variability:

Britney: Maybe, I would say the second dataset because, it's totally a guess but I guess, the eighteen is I guess is more like the median more. The repeating values are more outside. Seventy is further away from the eighteen and, oh well, ten to forty is kind of big jump. These values kind of less like in relation to each other.
Oguz: They are less in relation to each other.
Britney: I mean more spread apart. Like it goes, if it was in a dot diagram [might mean dot plot], all these numbers will be like on this end it kind of jumps like from eighteen to seventy either or side and this one lay down here [inaudible]. Well, actually maybe the first

> dataset would be more variable since it has more spread out. It has some datasets from the end, forty kind of in the middle, and seventy seven up here rather than the second one just having more like clustered on the lower end and the higher end.

Overall, both in interviews and in their responses to homework questions students tended to treat variability as the measure that gauges "how different the values are from each other." When this informal notion of variability was not applicable, however, students tended to view the concept of *variability* in different ways such as examining whether or not observations in a distribution are grouped in some certain areas.

Research Question 2:

In what ways, if any, does providing a context support or detract from students' reasoning about variability in the preceding situations?

Table 7 displays the frequency of students' choices for each version of the second homework question, those with and without context, with examples of student reasoning for each choice and each version. The differences in the distribution of percentages across the answer options suggested that although availability of context was not statistically significant in helping students to choose the correct option (39.4% for the version without context and 43.9% for the version with context), it did appear to change the incorrect choice made by students. In particular, students who saw the context version of the question were more likely than expected to choose that the distribution with more variability cannot be determined, while students who saw the version without context were more likely to choose option *Z* as the least variable.

Table 7

Percentages and Examples for Both Versions of Question 2

Option	Without Context		With Context	
	n %	Example	*n* %	Example
X	34 6.4	X has the least measure of variability because it has two occurrences of regions of data within the same range.	50 9.2	The items are more clustered and not as spaced out as the other options. Either her students went out 20 times a week or less than 3.
Y	209 39.4	In Y, more numbers were closer to the mean, 10, than the other sets.	238 43.9	Y is the answer because variability means spread and there is the least amount of spread in this data set because most of the data is concentrated in the middle near the mean.
Z	245 46.1	The data points that are depicted in the graph are equally spread out from each other, and thus they demonstrate a consistent, predictable pattern.	170 31.4	The students in the afternoon class went to eat at very constant intervals while the others were more sporadic.
Cannot be determ	43 8.1	Because they are both spread out	83 15.3	I said it cannot be determined because as we went over in the class and on the notes, each section has the same range of values.
Total	531		541	

Note. Correct choice was Y if students think of variability in terms of SD and variance.

Similar to the second question, the fifth homework question was asked both with and without a context. Coding 100 randomly selected responses suggested that students' decisions on the more variable dataset were similar in both versions of the question. The percentage of students who claimed that the second dataset was more variable was 81% for the group of students who took the question with the context, and 85% for the group of students who took it without a context. In addition to similar percentages in both versions of the question, fewer students included contextual information in their responses.

I investigated the influence of context on the interview participants' reasoning by offering a question without the context, observing the participants' approach to the question, and then providing the context and exploring whether they reconsidered their approaches. Analysis of the interview data did not suggest availability or lack of availability of context to have an impact on the interviewees' reasoning—especially for quantitative variables. In other words, availability of context did not suggest a strong clue or indication for interviewees to consider the variability of quantitative data in terms of how close, on average, observations were spread around the mean. For instance, in Task 2 of the first interview protocol, Chloe kept using her previous method of reasoning (which overlooked the mean and how data values were arranged in relation to the mean) even though she was able to discuss the centers of the distributions with the given context. Providing a context for the task seemed not to affect Chloe's approach. She slightly refined the context by suggesting "how many times people go downtown throughout the week" and claimed that the context for the data supported her reasoning. I further emphasized that all the distributions had the same average in order to see if she would take this new information into consideration in her reasoning. She claimed that either knowing the center of the distributions or having a context for them did not influence her preexisting thought process.

The interview participants occasionally used context to justify their informal notions of variability, as was illustrated in Karen's explanation of Task 4 of the first interview. She claimed that "[The first dataset] could be chocolate, vanilla, strawberry, but this [second dataset] is rainbow sugar, vanilla, and chocolate; you know it's so many other flavors." It should be noticed that the data on which Karen was working were quantitative, thus Karen's proposed scenario was inappropriate. In addition, she admitted that her reasoning was more of "thinking not numerically."

Additional Findings

Although there were no research questions posed to investigate students' reasoning about variability for categorical data, one of the homework questions (thereby, one of the interview tasks) required students to think about variability for categorical variables. In the eighth homework question (Figure 3), students were given two bar graphs of the same sample size, and asked to decide which distribution had greater variability. By including the question in students' homework assignments, I thought I might gather additional information about the informal notion that I targeted in Research Question 1.c. In addition, by asking this question I aimed to investigate the extent to which students conceptualized variability as a measure of how often the observations differ from one another.

The analysis of student responses to the question showed that 49.6% of the students selected the correct option that the Wednesday section was more variable, and 46.2% of the students selected that the Tuesday section was more variable. The other 4.2% of the students chose one of two options: the variability was the same for both sections or the variability of the distributions could not be determined. Observing that a large majority of the students chose either Tuesday or Wednesday as having more variability, for the rest of the analysis I investigated typical explanations they provided for these two options.

The bar graphs below show the frequency distribution of class year in two sections of a statistics lab.

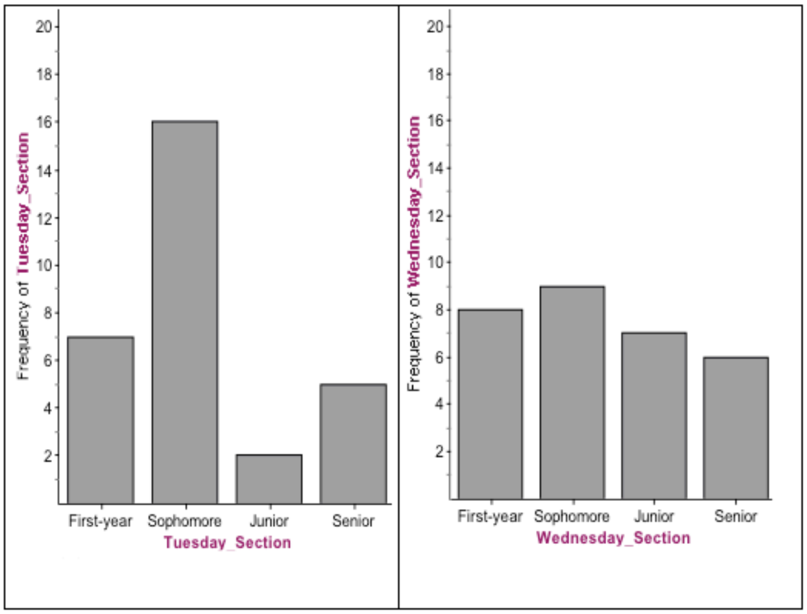


Figure 3. Homework question 8.

First, I examined the reasons for justifying the claim that the Wednesday section was more variable. An overwhelming majority of responses in this category pointed to the fact that the number of students was distributed across the categories almost equally in the Wednesday section; therefore, the Wednesday section would be more variable. Justifications were similar to the following student responses:

> Wednesday section has more variability because each year has about the same amount of students.
>
> Wednesday is the answer for that above question because the graph on the right [refers to the Wednesday section] displays more diversity in the amount of students that are present for all age groups. The one on the left [refers to the Tuesday section] displays more sophomores and less[*sic*] juniors. This takes away variability because the distribution of ages present on Tuesday is not displayed as evenly as in the right graph of Wednesday.
>
> In the Tuesday section there are 7 freshman (first-years), 16 sophomores, only 2 juniors, and 5 seniors. But in the Wednesday section, there are 8 freshman, 9 sophomores, 7 juniors, and 6 seniors. With that said, Wednesday's class has more variability, because there is almost the same amount of students from each class year in that section. For

> example, the probability of calling on a student in the Tuesday section and that student being a sophomore is relatively high because there are more sophomores in that section than any other class year. But the probability of calling on a student in the Wednesday section and them being a sophomore, freshman, junior, or senior is relatively going to be the same for each class year.

It is noteworthy that students were able to use their intuition to figure out what made a bar graph more variable. Only a few students, however, explicitly mentioned that the data represented in the question were categorical. In other words, the student responses did not clearly show that students recognized that the data represented in the question (through bar graphs) were categorical. On the other hand, some students used terms such as *mean, range, standard deviation, normal distribution, normal curve,* and a few more terms in order to justify their response. Such responses indicated that students might have treated the data shown as numerical. For instance, one student claimed that the Tuesday section was more variable, "because the standard deviation for the Tuesday section is much larger than the standard deviation for the Wednesday section."

In addition to treating the categorical data as numeric and inappropriately employing measures such as mean and standard deviation to bar graphs, the great majority of students concluded that the Tuesday section was more variable by observing the wider frequency differences for the Tuesday section categories. The following student responses depicted this widely observed approach:

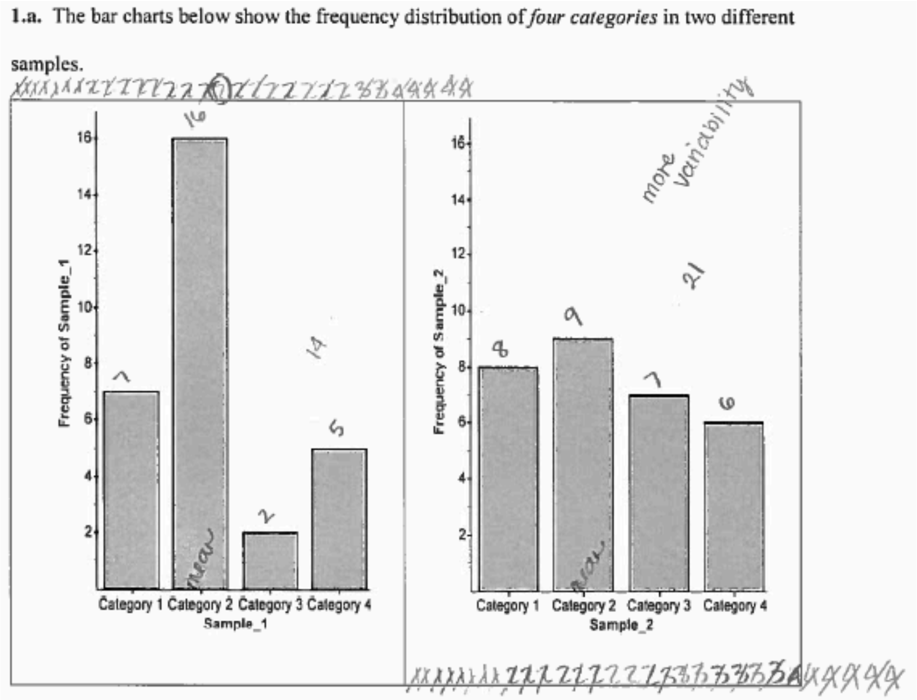
> The Wednesday section has an equal distribution of first year through senior students while for the Tuesday class, the amount of students based on class is different for each level, making it more varied.

> The Wednesday section has an equal distribution of first year through senior students while for the Tuesday class, the amount of students based on class is different for each level, making it more varied.

In conclusion, an analysis of student responses to this question indicated that most of the students who correctly identified the right answer used their intuition appropriately, whereas most of the students who chose the wrong answer used properties related to histograms and quantitative data in inappropriate ways.

In addition to the results from the analysis of student responses to the eighth homework question, findings of the interviews for Task 5 of the second interview, which was the same as homework question 8, suggested additional insight about students' ways of reasoning for categorical variability. In the following pages, I present a detailed report of each interview participant's approach to variability for categorical data as represented in bar graphs.

For Task 5 of the second interview, although Ocean clearly noted that "the bar graph tells the frequencies for specific categories" and the data under investigation in the task were categorical, such as "juice options" to choose from, she tried to calculate the mean values for each sample by using a method that was similar to computing median. As Figure 4 shows, Ocean applied an inappropriate method to compute the mean values for the categorical data. For both samples, she represented Category 1 with Number 1, Category 2 with Number 2, and so on, and she ranked them according to their frequencies. Then, she located the median value. According to this method, the mean for each sample was Category 2. Although such a strategy was inappropriate to follow for categorical data, I refrained from warning Ocean until she reached a conclusion based on her own calculations.



1.a. The bar charts below show the frequency distribution of *four categories* in two different samples.

Figure 4. Ocean’s method of calculating “mean” for categorical data.

Ocean claimed there were 14 data points “that were not included within the bar that includes the mean” in the first sample, and 21points that were not in the same category where mean was in the second graph. She further explained that “higher frequency that is different from the bin that includes the mean” would indicate the distribution to be more variable. Accordingly, she decided the first sample had a “smaller standard deviation” and therefore, was less variable.

Although Ocean decided that the second sample had larger variability, she kept thinking about the question, probably because she was not sure about the appropriateness of her conclusion. She even went back to her work in ranking the dot plots task and checked if her approach to this task was in accordance with her approach to the previous task. Upon observing that she was confused, I asked the second question of the task, assuming that the simplified version, which had only two categories for each sample, would help her to reason more easily. She approached the second question using the same method she generated for the first question.

Later, with some follow-up questions, Ocean recognized that calculating means for categorical data as she did in this task was inappropriate since the conventional meaning of mean is not valid for categorical data. On the other hand, she could not come up with another way that she might think of as appropriate for examining variability in bar graphs. Thus, she was unsure how to answer my question of "how do you interpret variability in categorical data?"

After she spent a few minutes working on the question, I provided the context to the data represented in bar graphs in the first question of the task, assuming that having a context might be helpful in her reasoning about categorical data. She immediately decided the Wednesday section (Sample 2) was more variable, but still had difficulty in explaining why she chose Sample 2. Then, I reiterated my question of how she would interpret variability in categorical data. She responded, "I would think you couldn't because unless you assign a variable number like one, two, three, or four to the categories … I don't think you could get numerical data or convert it to numerical information when you are given categorical chart." Ocean's reaction to this question and her following explanation seemed a noteworthy instance of how she attempted to come to a resolution to her dilemma:

Ocean: My explanation would be like with the mean and stuff, but you can't calculate a mean from these [points to the categories in the bar graphs. Because that was when I first think of variability. Now, I think of like average and stuff like that and the change the first definition that I gave you. Now when I think of variability I am picturing the average and—

Oguz: Maybe should you switch back to your previous?

Ocean: This is like *situational* [emphasis added] or like the t-shirt thing.

As the excerpt above shows, Ocean had difficulty in approaching variability in categorical data and bar graphs. Hearing that she remembered the t-shirt color example she provided to explain variability early in the first interview, I suggested that t-shirt colors—black and white—could represent categories in two different classes for the two samples given in the task. Using that context, Ocean correctly identified the more variable sample in both questions of the task.

Later, Ocean generated a method for assessing variability of a categorical variable based on the ratio between the frequencies of the categories. Ocean used the rule, if the ratio from one category to another is higher in one group, then that group had to have larger variability. She also said that the frequency differences among categories could be used to decide the bar graph with larger variability. She further claimed that the sample that has more categories should always be more variable. She elaborated on the issue by saying, "It has more variability because it has one more category, because you are giving more opportunity for different responses." Overall, from the variety of notions she provided, I was not sure if Ocean had adequate knowledge resources she could consistently use in the future when asked to reason about variability in categorical data.

As she spent more time thinking about the task, Ocean began to realize that variability might have different meanings in different "situations" (i.e., types of variables). As she claimed, "with the variability thing, it can alter based on the situation. You can't always go for average with especially with this situation where it is categorical data." It was clear from her expressions that Ocean started to recognize that the meaning of variability for categorical and numerical variables could be different.

Potential reasons for Ocean's difficulty in reasoning with categorical data seemed to stem from her lack of experience in thinking about variability for categorical data as well as her strong disposition on the meaning of variability for quantitative variables. This claim is further supported by her concern about categorical data: "Because you can't put a number on it and see where the average is at and how far each is away from the average and stuff like that." In other words, I suspected her "clustering around the mean" notion has dominated her ways of reasoning about variability and did not allow her to ease her thinking when it comes to categorical data. Another issue related to her difficulty with categorical variability might be because it is rarely addressed in typical statistics courses. Therefore, she might not have accumulated enough experience in investigating variability for categorical data.

In Task 5 of the second interview, Karen provided a description of the use of bar graphs that I found satisfactory: A bar graph basically presents "what categories have what frequency." Then, she started to work on the question by first suggesting a context for the data represented in given bar graphs:

> If we were thinking in terms of people, then a lot of people are in this category [points to the second category in the first bar graph], rather than these categories. Or, if I were to think of, it usually helps me to name the categories if it is really general like this it helps me to name the categories. So if these were countries and like if this one was Russia, I guess United States, Mexico, and Guatemala or something that would tell me that a lot of people live in this specific country [points to the second bar in the first bar graph] rather than these countries. So I feel like since more people live in this country [again, points to the second bar in the first bar graph] it doesn't have that much variety or variability if that makes sense. Whereas in this one [points to the second bar graph] they are close to each other, they are very close to each other. There is pretty … they are close to each other. There is pretty much the same amount in each category, yeah people I guess. So as far as this goes I would choose B [as the more variable sample]. Sample two has more variability.

It was obvious that Karen had an idea of what she needed to do in order to reason about variability for categorical variables. Therefore, instead of asking for more detail on how she worked on the question, I asked her to explain her interpretation of variability. She said, "As far as categorical, variability means variety, variability means that the frequencies in each category to be similar." Karen's work on the question and the her clear explanation suggested that she was able to describe the concept of *variability* for categorical data and employ her description when it comes to reasoning about categorical variables as presented in bar graphs. Upon my question on whether variability is same for categorical and quantitative variables, she claimed they were "similar but organized differently." She further explained that variability was based on "frequency" in bar graphs and she did not need to check "the range, spread, or any of that" when reasoning about variability in categorical data.

In the Aquarium and Zoo question of the task, Karen provided explanations that were again satisfactorily indicative of her understanding of variability for categorical data. Some of these explanations were, "so much more students lean toward aquarium so not much variability" and "Students [refers to the seventh grade] are even in both sides and it makes it more variable." Accordingly, she suggested that she "would choose grade seven ultimately because it's like pretty much the same frequency" across categories. Karen's explanations suggested that she focused mostly on the (differences of) frequencies of categories in bar graphs in her reasoning about categorical data. Overall, as far as categorical data—thus, bar graphs—were concerned, Karen regarded variability as *variety* and made decisions based on this very notion.

In Task 5, Chloe noted that all the frequencies were very similar in the second sample; thus the first sample had to be more variable. The expression that she used to explain her reasoning included her saying that the first bar graph had an "up-down thing" of the categories. Overall, it seemed that Chloe strived to formulate her reasoning based on the frequency differences across the categories. Smaller frequency differences among categories of a bar graph generally suggest larger variability, which was also valid in this question. However, Chloe's response suggested that she saw things the opposite way.

Next, I provided the version of the question that included context. Chloe discovered that the Sophomore Category constituted the majority of the frequencies in the Sample 1, but frequencies were distributed among the frequencies in approximately the same amounts in

Sample 2. Observing that the first sample had the “overwhelming amount of sophomores, not many juniors, some freshmen, and some seniors, ” she decided that the first sample was less variable. Accordingly, she said that none of the categories were “over powering the other” in the Sample 2.

Chloe employed a different strategy in the third question; therefore, although I was expecting her to conclude that seventh grade was the most variable (since both aquarium and zoo categories had equal frequencies), she decided that sixth grade was the most variable. She explained her thought process in the following excerpt:

> Because in the seventh grade … not variability at all, same amount of kids, fifty fifty, it’s halved. Eight grade … overwhelming majority, eighty kids for the aquarium, twenty for the zoo, it’s a higher range but it’s not like it’s varied because that’s like out of ten kids eight of them say this, here is eight and here is two, it’s not varied at all, most of it is in that. This one, grade nine, it’s a little better, but it’s still so far away like, seventy to thirty, that’s like, thirty, seventy, and you know, I am looking at, like, that in my head. And grade six, they are a little bit closer to each other but enough part apart. You are definitely like say you are voting like and this is saying, this has majority to win [points to the higher bar in Grade 9], this has majority to win [points to the higher bar in Grade 8], this has to have a even break [points to the bars in Grade 7], this [points to the bars in Grade 6] okay, okay forty to sixty almost half and half but not half and half so there is a good amount of kids that like this one versus the other one but nothing is overwhelming the other, so it’s sixth grade.

Chloe’s explanation was surprising. For Chloe, a “fifty-fifty” distribution of frequencies in double-bar graphs was insufficient to claim that a distribution was the most variable. She claimed, “Equality of frequencies across categories indicates variability but not much.” As it was evident from her explanation, Chloe required a small amount of frequency differences across categories between frequencies in order to conclude that a bar graph was more variable. Even after my follow-up questions and reminding her of her way of thinking for the previous question, Chloe maintained her conclusion that the sixth grade was the most variable class because the frequencies in sixth grade’s data were very close but still were not equal.

Overall, Chloe suggested, “Determining variability in categorical is a little bit different than quantitative one.” For quantitative data, she suggested range and availability of various numbers (“hitting more numbers” in her own words) as the two most crucial criteria. For categorical data, she claimed that categories should have similar amount of frequencies but not equal in order to conclude that a bar graph represents a more variable dataset.

When Britney started to work on Task 5, first I wanted to understand how she could extract the information represented in bar graphs. She suggested that bar graphs presented “the frequency to the categories.” When I asked which sample had more variability in terms of the categories, Britney claimed that the second sample was more variable. She justified her conclusion by claiming, “frequencies being similar throughout causes more variability.” She elaborated her thinking saying, “They are all equally affecting on the decision because this is all pretty much high frequency rather than having like a small bar down here that does not affect much.”

In addition, Britney noted, “The bins [refers to bars in each category] are also different from each other.” This additional requirement noted by Britney was noteworthy especially because Chloe also mentioned it as a criterion in her work with bar graphs that came next. In brief, both Chloe and Britney claimed that even though an approximately equal distribution of frequencies across categories suggested that the distribution was more variable, the frequencies had to have slightly different frequencies (although Britney did not apply this criterion when she worked on the last question of the task).

When I asked the second question of the task in which there were only two categories in each sample, Britney again concluded that the second sample was more variable. The rationale she provided was as, “Nothing is more favorite, nothing is more prominent over the other one.” The following excerpt provides a more comprehensive account of her reasoning:

> If you had fourteen people go to this event and then sixteen people go to this event, it’s like frequency close for each. It’s a little higher but it’s still close, really nothing is more favorite, nothing is more prominent over the other one so it has more variability because it’s even and close to been even. And this one [refers to the first sample] in context a lot more people go to this event rather than this one. So the variability of people going to events is not that much because you can tell that most people go to this one.

Britney’s explanation suggested that she was able to think about variability for categorical variables and articulate her thinking upon my questions. When I asked if the ways of reasoning for variability were different for quantitative and categorical data, she said that it was “Still the same reasoning but just looks different.” Her response suggested that she was not able to recognize that variability is interpreted differently in categorical and quantitative variables.

For the last question of the task, in which four different school grade’s preference were presented in double bar graphs, Britney maintained her approach and said that in the more variable class (which was the seventh grade), the options were “equal to each other” and “Nothing is favored, so nothing is prominent over the other.” She also added, “if there is a lot of variability you cannot make a decision easily.”

Overall, student responses to the eighth homework question suggested that almost half of the students were able to reach a correct conclusion by applying their intuitive ideas of variability for the categorical data and bar graphs. The other half of the students, however, reached an incorrect conclusion, probably because they treated bar graphs as if they were histograms and then applied the properties related to histograms and quantitative data in inappropriate ways to the bar graphs. Students’ performance during the interviews suggested that (a) one of the interview participants was not able to reason about variability for categorical data although she was competent in reasoning about variability for quantitative data and (b) other participants were able to reason about variability for categorical data although they had serious difficulties in reasoning about variability for quantitative data. The findings suggest that learning to reason about variability for different types of data might be independent from each other. Lastly, interview participants found the contextual information helpful in reasoning about variability for categorical data.

CHAPTER 5

DISCUSSION

In this study, I examined the assumption that students' reasoning about variability would be influenced by the prevalent characteristics of the distributions and datasets on which students were asked to work. By collecting and analyzing homework and interview data from undergraduate students enrolled in an introductory statistics course, I was able to frame students' particular ways of reasoning about variability. In this chapter, first, I summarize the findings reported in Chapter 4. Next, I describe the limitations of the study. Finally, I discuss implications for instruction and further research.

Summary and Discussion of Results

In the last chapter, I presented findings from homework questions and interviews. In this section, first I summarize findings for each research question of the study. Next, I provide a discussion of the results.

Research Question 1.a

The results suggested that many students focused on the differences among the ranges of the data when comparing variability across different distributions. When given distributions had equal ranges, these students suggested different approaches for the concept *variability*. When given raw data or dot pots with only a few values, a considerable group of students claimed that the distribution with values distinct from each other had more variability. Acoordingly, if some or most of the values in a distribution were grouped together, students indicated that the distribution was less variable because the clustering of values meant that the values were more "similar to each other." Note that students often lacked an integral aspect of the formal definition of variability, which focuses on the relative position of the grouping of values with respect to the center of the distribution.

When data were presented in dot plots and histograms, many students equated the concept *variability* with the differences in frequencies of observations in these graphical displays. For example, some students claimed that approximately uniform distributions should be less variable than normal distributions because the frequencies of bins for the first would be very similar to each other, whereas the frequencies in normal distributions would fluctuate more, thereby exhibiting more variability. There were, however, some students who claimed that approximately uniform distributions should be more variable than normal distributions or distributions with different shapes. These students often justified their thinking based on the claim that in uniform distributions each value (such as the bins of a histogram or horizontal values in a dot plot) is equally likely to occur, whereas bins or horizontal values have different probabilities of occurance in non-uniform distributions; therefore, uniform distributions are supposed to be more variable. As is clear from this way of thinking, these students tend to think about variability in a way that is more applicable to categorical data.

Research Question 1.b

Analyzing both homework and interview data suggested that when students were assessing the variability of a distribution, they did not focus on the notions of presence of individual values in a distribution. In other words, students did not claim a distribution was less or more variable based solely on the availability of extreme values or outliers. Many students noticed those values when responding to the homework questions or explaining their reasoning in interviews, but used them as a secondary phenomenon to consider when explaining their reasoning about variability. For example, three of the interview participants used extreme values and outliers to support their reasoning based on the notions of *range* and *variety of data values*. Similarly, almost half of the students who noticed outliers or extreme values in a distribution used those individual values to justify their conclusions about the shape of a distribution (e.g., to justify the claim that a distribution was *skewed*). The other half of the students claimed that the presence of outliers indicated a bigger variability for various reasons (e.g., outliers suggest that the range is greater and outliers mean a greater variety of observations in a distribution). Overall, the results of the analysis of homework and interview data collectively suggested that availability or unavailability of outliers did influence students' reasoning about variability extensively. Many students tended to use their preexisting notions and approaches to assessing variability in similar ways both with and without the presence of extreme values.

Research Question 1.c

The treatment of variability as *how different the values are from each other* was the overarching theme when each interview participant described variability at the beginning of the interviews. Three of the interview participants also used this notion throughout the interviews when a task permitted its application. When this notion was not applicable, because of the characteristic of the question or other features of the tasks that overshadowed the use of the notion, the participants classified the distributions as being equally variable. Similarly, in their responses to homework questions many students were inclined to treat variability as a measure that gauges the *variety data values* if one of the datasets had more varied data values than the other. Unfortunately, the results of the analysis of homework data were inconclusive in suggesting insights about the research question because the homework questions did not provide an opportunity to compare varaibility when the datasets or distributions to be compared had approximately the same number of different values.

Research Question 2

Results of the student responses to homework questions suggested that although availability of context was not statistically significant in helping students to choose the correct option, it did appear to change the incorrect choice made by students. Analysis of homework data showed that students' use of context-related language in their explanations about the variability in the given datasets were not common. Similarly, the analysis of the interview data did not suggest that availability or lack of availability of context had a positive impact on the interviewees' reasoning—especially for quantitative variables. In other words, availability of context did not appear to provide clues or indications to interviewees that addressing variability of quantitative data requires an assessment of how close, on average, observations are spread from the mean.

Discussion of Results

Overall, analysis of interview and homework data suggested a relationship between students' reasoning and the prevalent characteristics of the distributions on which students were asked to work. For example, students addressed the shape of a distribution when given skewed distributions more frequently than they addressed shape and variability of distributions with other shapes.

The analysis of the student responses to homework questions also suggested that students generally had a rule-based approach to statistical questions in their response to homework questions. For instance, for the first question in Appendix B, instead of focusing on what skewness could mean in the context of coffee consumption, students usually applied simple rules of statistics related to skewed distributions. For example, many students stated the relationship between a skewed distribution and its implication on whether mean or median of the distribution is larger, giving a response such as, “The distribution is skewed towards the right. That means that the mean is greater than the median.”

The homework data were occasionally insufficient to answer the research questions, especially for some of the questions asked in homework assignments. As reported in Chapter 4, students often did not mention variability of a distribution if not specifically asked to do so. Even when students were asked to describe variability in a given distribution, they rarely discussed variability in detail. The majority of the students seemed to overlook the variability of the distribution in their responses, which was consistent with previous studies (e.g., Cooper & Shore, 2008; Kaplan, Lyford, Jennings, & Gabrosek, in press; Meletiou & Lee, 2002) that students often fail to attend to variability when describing histograms.

The results of the study fell short in answering some of the research questions also because the data collection means were not detailed and comprehensive enough to extract rich data to answer these research questions. For example, in the second research question, I aimed to investigate the role of context in students’ reasoning about variability, but providing contexts did not appear to suggest to students the need to use the mean of a distribution as the important theme related to variability. The student responses in both homework and interviews did not suggest that context was useful for students to recognize the type of data represented in the graphs and its implications in terms of thinking about variability. The finding, however, was contradictory to my assumption that context would help the student think the bins were quantitative and not categories.

Availability of context helped all four interviewed students in their approach to reasoning about variability for categorical data and articulating their reasoning behind it. In that sense, the finding supports the claim made by other researchers (e.g., English, 2012; Pfannkuch, 2011) that context can be an aid or an obstacle in students' reasoning. The finding is noteworthy because students' fluency in thinking about variability of categorical data might help them recognize the different types of data (such as categorical versus quantitative and univariate or bivariate quantitative).

Some of the findings that emerge from the present study are similar to those in previous studies. For example, the findings agree with Garfield et al. (2007) that students usually maintain their thinking about variability as "overall spread and differences in data values (e.g., not all values are the same)" (p. 142). Ways of thinking, such as (a) smaller range values mean less variability because the values will be similar and (b) when values are close together they are more similar hence less variable, were commonly observed among students. All in all, approaching variability in these ways was not totally unhelpful; these notions could, in fact, offer a useful intuitive basis for understanding variability as a measure of how data values cluster around the mean, although I make no claim that most of the college students were prone to this way of thinking.

Findings of the study also contrast with some of the previous research findings. The research literature, for instance, has suggested that students tend to focus on individual values and especially extreme values when they reason about variability, which was not common among the interview participants of the study or the students who submitted their responses to homework questions. One possible reason that could explain the disagreement between the findings of the present study and previous studies might be the use of a few questions that targeted this phenomenon in the present study. Including more homework questions and interview tasks could substantially increase the overall richness of the data in answering the research question related to extreme values.

The study also added more detailed information about the knowledge resources that undergraduate college students employ when reasoning about variability and determining more or less variable distributions, the characteristics of given distributions that were influential in their approach to variability, and the role of context. In addition, students' tendency to treat variability for quantitative data in a way that is more suitable for categorical data might be because of the everyday meaning of *variability, vary,* and *variation*. As also supported by the first descriptions students provided in the beginning of the interviews, students usually think about variability in its colloquial meaning. Overall, the study contributes to the statistics education research literature by critically investigating the interaction between the content students were asked to work on and their approach to variability.

Limitations

The interviews were conducted with a small group of students a few times during the study. Ideally, more interviews with more students would provide a better depiction of students' reasoning. In addition, results of the interviews were contingent upon the type of questions in the interview protocols and the students' effort when responding to these questions. Although the interview tasks in my study had many follow-up questions, the interview data could be limited in terms of laying out students' individual thinking mechanisms. It might have needed to use a more varied and tailored set of questions for each interviewee based on their unique ways of thinking.

The study included students' responses to eight homework questions collected from all registered students in the course. Ideally, more questions would be asked so that a more thorough depiction of students' reasoning could be achieved. There were some restrictions in terms of putting more questions into students' homework assignments for research purposes because the course already had several required assignments every semester. In addition, the results of the study were highly influenced by the extent to which students took the questions seriously or attempted to answer to the best of their abilities. Students often explained their thinking briefly in their responses to homework questions, which resulted in difficulties for deducing students' exact reasoning about variability.

In the study, I focused more on identifying students' informal notions of variability, which might have caused me to neglect other important issues. The researcher's focus might also influence the selection of data collection such as including only some certain types of statistical tasks but missing other important aspects of students' reasoning about variability. Therefore, interpreting the results of both interview and student responses to homework data has constraints. First of all, one can claim that the situation in which students were asked to reason about variability in the study had a major impact on what type of responses student could possibly provide. In other words, it is reasonable to assert that the design of the study overall and the questions asked specifically in the interviews and homework questions might have directed students to think about variability in certain ways. For example, different types of data representations and their prominent characteristics might have triggered certain types of reasoning. As I discussed when I proposed the Theoretical Framework, novice knowledge is dependent on the circumstances and conditions under which students are asked to reason. In the study, the characteristics of the datasets, distributions, and even the small differences (that even I could not be aware of) might have led interviewees to think in certain ways. Therefore, the same students could have shown different reasoning mechanisms if the interview tasks had been organized differently.

Lastly, the study relied on qualitative data analysis, which introduces a limitation called lack of internal consistency. Most quantitative techniques will result in the same results if the data remain the same. Qualitative data analysis, however, depends heavily on the researcher. Thus, the results of the analysis may differ across different researchers analyzing the data. By recruiting two or more researchers to work on the data, inter-rater reliability could be achieved. Although a capable colleague occasionally examined my coding of students' responses to homework questions, the overall coding process was essentially based on my own coding; thus, inter-rater reliability was not achieved for the analysis and results of students' responses to homework questions in the present study. Similarly, I did not aim to achieve an inter-rater reliability for the analysis of interview data.

Implications for Further Research

Many students were very brief when explaining their thinking in their responses to homework questions, which resulted in difficulties for deducing students' exact reasoning about variability. Lack of detail in students' explanations suggested that using homework data this way may be an obstacle to sound research findings, especially if the goal of a research study is to achieve detailed and in-depth data. The problem with the homework data also raises the issues about students' motivation toward answering homework questions. In my case, students probably knew they would get points automatically for whatever explanations they submitted in their response, which might have reduced their effort and motivation to answer homework questions to the best of their abilities. Therefore, researchers who want to collect large-scale data may need to consider the motivation aspect, seek other ways to collect data from students, and be cautious in interpreting students' knowledge in answering constructed-response questions that are not subject to grading. Accordingly, researchers should be careful in generalizing findings that they extract from similar homework data since these findings have a potential to underrepresent overall student performance.

Integrating the knowledge-in-pieces epistemological perspective framework into my study was helpful in making sense of the interview findings. Three of the interview participants presented a fragmented understanding of variability, as their approaches to variability were inconsistent across the tasks. In other words, those participants went back and forth between various approaches in different situations. As hypothesized before the present study was carried out, if students were given raw data with some data values repeating (i.e., having the same numerical values more than once in a distribution), they interpreted variability as *difference* and compared variability across distributions according to the *difference* notion. In this specific case, many students found comparing ranges across datasets to be inconvenient and inappropriate. According to the knowledge-in-pieces epistemological perspective, however, such a student behavior is expected because undergraduate students can be regarded as novices in reasoning about variability, thus their understanding and knowledge were heavily influenced by the characteristics of the tasks that they were asked to work with. Accordingly, I agree with Bakker's (2004) suggestion that both instruction and research on students' understanding of statistical ideas need to pay attention to providing experiences with a variety of distributional shapes, contexts, and variability. In conclusion, as students develop an understanding of variability, they should start to recognize that one or two characteristics of a distribution, such as having a larger range or extreme values, is insufficient to claim that the distribution has a larger variability.

Wagner (2006) suggested that different *situations* have their own affordances and facilities with respect to students' reasoning about a concept. Similarly, Meletiou and Lee (2002) claimed that students' reasoning about variability is heavily reliant upon both the particularities of the task explored, which refers to situations and the contexts in which the tasks are situated. Use of the knowledge-in-pieces perspective was instrumental in investigating ways in which the characteristics of distribution, graphical representation, and statistical context cue different ways of reasoning, such as the one I observed in student responses to the second homework question. As extensively discussed in Chapter 4, students' explanations were different in nature than the explanations they provided to the more typical statistical questions such as describing histograms or comparing variability across dot plots with more common shapes.

The results of the interviews provided some new evidence for possible learning trajectories for variability, which could be studied further in the future. According to the interview results, students who had a solid quantitative understanding of variability in quantitative data had difficulty understanding variability in categorical data. On the other hand, students who did not have a robust understanding of variability for quantitative variables were, in general, able to reason about variability in categorical data. These findings could be used further in examining whether students tend to think about variability in two different ways, *how often* the observations differ from one another and *how much* the observations differ from the mean, or if students hold different ways of reasoning about variability that lie between these two fundamental ways of reasoning about variability.

In addition to its implications for possible learning trajectories, the study presented students' particular difficulties with working with variability. The students who recognized their difficulties with variability mentioned them during the interviews. Those insights give important evidence about how students learn statistical concepts. For example, one of the interview participants said she had difficulty conceptualizing variability when comparing two samples of categorical data with the same number of categories in each sample. The student said it was easier for her to compare two samples and decide on the sample with smaller variability if one of the samples has fewer categories. This observation and others present important information on how to teach variability for categorical data. Overall, future studies could use the results of the present study in investigating students' affordances and difficulties with the statistical concept of variability.

Results of the study suggested that students use the appropriate statistical terminology sparsely when asked to explain their thinking about statistical variability. The study clearly showed that interviewed students tended to use the colloquial meaning of *variability, variation,* and *vary* and apply them to the core statistical situations haphazardly. In addition, it was common from both data resources of the study that students used *range* and *spread*, often without giving much attention to their specific meaning in statistics. For instance, take the use of *range* by students. The statistical term *range* is calculated for quantitative variables, and it is based on the minimum and maximum quantitative data values, not based on the minimum and maximum frequencies for categories or bins. In contrast, the interview participants often used *range* to refer to the "range of frequencies." Therefore, instructors and researchers should focus on students' use of statistical language, and assess whether students use the proper statistical meanings of the terms. As conceptualized as *lexical ambiguity* (e.g., Kaplan et al., 2010; Kaplan, Rogness, & Fisher, 2014) the term variability requires explicit attention. Future studies could be conducted to enable researchers and instructors to exploit students' preexisting notions of the word *variability* in introducing the statistical term. These studies could first examine whether and to what extent students distinguish the term’s meaning in statistics and outside the world of statistics, and then suggest instructional interventions that could help students conceptualize statistical variability in the ways curricular documents (e.g., ASA, 2005; Franklin et al., 2007) and pertinent research literature promote.

REFERENCES

American Statistical Association (ASA). (2005). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author.

Arnold, P. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data* (Doctoral dissertation). The University of Auckland, New Zealand.

Arnold, P. M., & Pfannkuch, M. (2014). Describing distributions. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)* (pp. 1–6). Voorburg, The Netherlands: International Statistical Institute.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83.

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill, & M. Camden (Eds.), *Curricular development in statistics education: International Association for Statistical Education 2004 Roundtable* (pp. 163–173). Voorburg, The Netherlands: International Statistical Institute.

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147–168). Dordrecht, The Netherlands: Kluwer.

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

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*(1), 35–65. doi:10.1023/a:1013809201228

Bock, D. E., Velleman, P. F., & De Veaux, R. D. (2012). *Stats: Modeling the world–AP edition* (4th ed.). Boston, MA: Pearson.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Canada, D. L. (2004). *Elementary preservice teachers' conceptions of variation* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (Order No. 3150610)

Ciancetta, M. A. (2007). *Statistics students reasoning when comparing distributions of data* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (Order No. 3294660)

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

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1*(1), 5–43.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. doi:10.3102/0013189x032001009

Confrey, J., & Makar, K. (2002). Developing secondary teachers' statistical inquiry through immersion in high-stakes accountability data. In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), *Proceedings of the 24th Annual*

Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (Vol. 3, pp. 1267–1279). Columbus, OH: ERIC.

Cooper, L. L., & Shore, F. S. (2008). Students’ misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education, 16*(2), 1–13.

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty with reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy*. Auckland, New Zealand: University of Auckland. Retrieved from https://apps3.cehd.umn.edu/artist/articles/srtl4_artist.pdf

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.

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

delMas, R., & Liu, Y. (2007). Students' conceptual understanding of the standard deviation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 87–116). Mahwah, NJ: Erlbaum.

diSessa, A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Hillsdale, NJ: Erlbaum.

diSessa, A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*, 105–225.

diSessa, A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education, 20*(10), 1155–1191.

diSessa, A. A., Sherin, B. L., & Levin, M. (2016). Knowledge analysis: An introduction. In A. A. diSessa, M. Levin & N. J. S. Brown (Eds.), *Knowledge and interaction: A synthetic agenda for the learning sciences* (pp. 30–71). New York, NY: Routledge.

Elby, A. (2000). What students' learning of representations tells us about constructivism. *Journal of Mathematical Behavior, 19*(4), 481–502. doi:http://dx.doi.org/10.1016/S0732-3123(01)00054-2

English, L. D. (2012). Data modeling with first-grade students. *Educational Studies in Mathematics, 81*, 15–30. doi:10.1007/s10649-011-9377-3

English, L. D., & Watson, J. M. (2015). Exploring variation in measurement as a foundation for statistical thinking in the elementary school. *International Journal of STEM Education, 2*(1), 1–20. doi:10.1186/s40594-015-0016-x

Franklin, C. (2013). Common core state standards and the future of teacher preparation in statistics. *The Mathematics Educator, 22*(3), 3–10.

Franklin, C., Bargagliotti, A. E., Case, C. A., Kader, G. D., Scheaffer, R. L., & Spangler, D. A. (2015). *The statistics education of teachers*. Alexandria, VA: American Statistical Association.

Franklin, C., Kader, G. (2006). A sequence of activities for developing statistical concepts. *Statistics Teacher Network*, *68*, 1–11.

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 preK-12 curriculum framework.* Alexandria, VA: American Statistical Association.

Friel, S. N., Curcio, F. R., and Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124–158.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*, 1–51.

Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 47–78). Boston, MA: Kluwer Academic Publishers.

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92–99.

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

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notion of variability to develop an understanding of formal measures of variability. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 117–147). Mahwah, NJ: Erlbaum.

Garfield, J., & Gal, I. (1999). Teaching and assessing statistical reasoning. In L. Stiff (Ed.), *Developing mathematical reasoning in grades K-12* (pp. 207–219). Reston, VA: National Council of Teachers of Mathematics.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. London, UK: Weidenfeld & Nicholson.

Goldin, G. A. (1997). Chapter 4: Observing mathematical problem solving through task-based interviews. *Journal for Research in Mathematics Education Monograph, 9*, 40–62. doi:10.2307/749946

Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 517–545). Mahwah, NJ: Erlbaum.

Gould, R., & Ryan, C. (2014). *Instructor's edition: Introductory statistics Exploring the world through data*. Boston, MA: Pearson.

Graham, A. T., Pfannkuch, M., & Thomas, M. O. J. (2009). Versatile thinking and the learning of statistical concepts. *ZDM, 41*, 681–695. doi:10.1007/s11858-009-0210-8

Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17–41.

Humphrey, P. B., Sharon, T., & Mittag, K. C. (2013). Developing consistency in the terminology and display of bar graphs and histograms. *Teaching Statistics, 36*(3), 70–75.

Izsák, A. (2005). Learning to frame research in mathematics education. *The Mathematics Educator, 15*(2), 2–7.

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014) Establishing the validity of the locus assessment through an evidenced-centered design approach. In K. Makar, B. de Sousa, & R. Gould (Eds.) *Proceedings of the Ninth International Conference on Teaching Statistics* (pp. 1–6). Voorburg, The Netherlands: International Statistical Institute.

Jacobson, E., & Izsák, A. (2014). Using coordination classes to analyze preservice middle-grades teachers' difficulties in determining direct proportion relationships. In J. Lo, K. R. Leatham, & L. R. Van Zoest (Eds.), *Research Trends in Mathematics Teacher Education* (pp. 47–65). New York, NY: Springer.

Jennings, J. K. (2014). *Calibrating test item banks for an introductory statistics course* (Master's thesis). Available at https://getd.libs.uga.edu/pdfs/jennings_jeremy_k_201405_ms.pdf

Jones, D. L., & Scariano, S. M. (2014). Measuring the variability of data from other values in the set. *Teaching Statistics, 36*(3), 93–96.

Jones, G. A., Langrall, C. W., Thornton, C. A., Mooney, E. S., Wares, A., Jones, M. R., . . . Nisbet, S. (2001). Using students' statistical thinking to inform instruction. *Journal of Mathematical Behavior, 20*(1), 109–144.

Kader, G., & Jacobbe, T. (2013). *Developing essential understanding of statistics, Grades 6–8*. Reston, VA: National Council of Teachers of Mathematics.

Kader, G., & Mamer, J. (2008). Statistics in the middle grades: Understanding center and spread. *Mathematics Teaching in the Middle School, 14*(1), 38–43.

Kader, G., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education, 15*(2), 1–17.

Kaplan, J. J., Fisher, D. & Rogness, N. (2010). Lexical ambiguity in statistics: How students use and define the words: *association, average, confidence, random* and *spread*. *Journal of Statistics Education, 18*(2). Retrieved from http://www.amstat.org/publications/jse

Kaplan, J. J., Gabrosek, J. G., Curtiss, P., & Malone, C. (2014). Investigating student understanding of histograms. *Journal of Statistics Education, 22*(2), 1–30.

Kaplan, J. J., Lyford, A., Jennings, J. K., & Gabrosek, J. G. (in press). Investigating student descriptions of histograms.

Kaplan, J. J., Rogness, N., & Fisher, D. (2014). Exploiting lexical ambiguity to help students understand the meaning of random. *Statistics Education Research Journal, 13*(1), 9–24. Retrieved from http://iase-web.org/Publications.php?p=SERJ

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.) *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259–289. doi:10.2307/749741

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics, 88*(3), 305–325. doi:10.1007/s10649-013-9529-8

Lann, A., & Falk, R. (2003). What are the clues for intuitive assessment of variability? In C. Lee (Ed.), *Proceedings of the Third International Research Forum on Statistical Reasoning, Thinking, and Literacy* (pp. 1–24). Mount Pleasant, MI: Central Michigan University.

Lather, P. (2004). Critical inquiry in qualitative research: Feminist and poststructural perspectives: Science "after truth". In K. DeMarrais & S. D. Lapan (Eds.), *Foundations for research: Methods of inquiry in education and the social sciences* (pp. 203–216). Mahwah, NJ: Erlbaum.

Leavy, A. M., & Middleton, J. A. (2011). Elementary and middle grade students' constructions of typicality. *Journal of Mathematical Behavior, 30*(3), 235–254. doi:10.1016/j.jmathb.2011.03.001

Lehrer, R., Kim, M.-J., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM, 43*, 723–736. doi:10.1007/s11858-011-0347-0

Lehrer, R., & Schauble, L. (2002). Distribution: A resource for understanding error and natural variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics* (pp. 1–6). Voorburg, The Netherlands: International Statistical Institute.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635–679.

Leinhardt, G., & Larreamendy-Joerns, J. (2007). Discussion of Part I: Variation in the meaning and learning of variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 177–190). Mahwah, NJ: Erlbaum.

Loosen, F., Lioen, M., & Lacante, M. (1985). The standard deviation: Some drawbacks of an intuitive approach. *Teaching Statistics, 7*(1), 2–5.

Makar, K. (2016). Developing young children's emergent inferential practices in statistics. *Mathematical Thinking and Learning, 18*(1), 1–24. doi:10.1080/10986065.2016.1107820

Makar, K., & Confrey, J. (2005). “Variation-talk”: Articulating meaning in statistics. *Statistics Education Research Journal, 4*(1), 27–54.

Maxwell, J. A. (2013). *Qualitative research design: An interactive approach* (3rd ed.). Thousand Oaks, CA: Sage.

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics* [CDROM]. Voorburg, The Netherlands: International Statistical Institute.

Meletiou-Mavrotheris, M., & Lee, C. (2010). Investigating college-level introductory statistics students' prior knowledge of graphing. *Canadian Journal of Science, Mathematics and Techology Education, 10*(4), 339–355. doi:10.1080/14926156.2010.524964

Moore, D. S. (1988). Should mathematicians teach statistics? *College Mathematics Journal, 19*(1), 3–7.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). Washington, DC: The National Academies.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices, Council of Chief State School Officers (CCSSI). (2010). *Common core state standards for mathematics*. Washington, DC: Author.

Noll, J., & Shaughnessy, J. M. (2012). Aspects of students' reasoning about variation in empirical sampling distributions. *Journal for Research in Mathematics Education, 43*(5), 509–556.

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Meanings of average and variation in nursing practice. *Educational Studies in Mathematics, 40*, 25–51.

Pearl, D. K., Garfield, J. B., delMas, R., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). *Connecting research to practice in a culture of assessment for introductory college-level statistics.* Retrieved from http://www.causeweb.org/research/guidelines/ResearchReport_Dec_2012.pdf

Peck, R., Gould, R., & Miller, S. J. (2013). *Developing essential understanding of statistics, Grades 9–12*. Reston, VA: National Council of Teachers of Mathematics.

Perry, M., & Kader, G. (2005). Variation as unalikeability. *Teaching Statistics, 27*(2), 58–60.

Peters, S. A. (2010). Engaging with the art and science of statistics. *Mathematics Teacher, 103*(7), 496–503.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning, 13*(1–2), 27–46. doi:10.1080/10986065.2011.538302

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

Pingel, L. A. (1993). Variability-Does the standard deviation always measure it adequately? *Teaching Statistics, 15*(3), 70–71.

Pirie, S. E. B. (1996). What are data? An exploration of the use of video recording as a data gathering tool in the mathematics classroom. In I. E. Jakuboski, D. Watkins, & H. Biske (Eds.), *Proceedings of the Sixteenth Annual Meeting of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 553–559). Columbus, OH: ERIC.

Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *Journal of Mathematical Behavior, 22*, 405–435. doi:10.1016/j.jmathb.2003.09.002

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

Scheaffer, R. L. (2000). Statistics for a new century. In M. J. Burke & F. R. Curcio (Eds.), *Learning mathematics for a new century* (pp. 158–173). Reston, VA: National Council of Teachers of Mathematics.

Schoenfeld, A. H., Smith, J. P., & Arcavi, A. (1993). Learning: The microgenetic analysis of one student's evolving understanding of a complex subject matter domain. *Advances in Instructional Psychology*, *4*, 55–175.

Shaughnessy, J. M. (2006). Research on students' understanding of some big concepts. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance: 68th yearbook* (pp. 77–98). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2, pp. 957–1009). Charlotte, NC: Information Age.

Strauss, A., & Corbin, J. (1998). Basics of qualitative research: Procedures and techniques for developing grounded theory. Thousand Oaks, CA: Sage.

Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. A. Duschl & R. J. Hamilton (Eds.), *Philosophy of science, cognitive psychology, and educational theory and practice* (pp. 147–176). Albany, NY: State University of New York Press.

Tabor, J., & Franklin, C. (2013). *Statistical reasoning in sports*. New York, NY: Freeman.

University of Georgia. (2015). *Fall 2015 UGA Bulletin*. Retrieved from http://www.bulletin.uga.edu/index.aspx

Vermette, S., & Gattuso, L. (2014). High school teachers' pedagogical content knowledge of variability. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics* (pp. 1–6). Voorburg, The Netherlands: International Statistical Institute.

Wagner, J. F. (2006). Transfer in pieces. *Cognition and Instruction, 24*(1), 1–71. doi:10.1207/s1532690xci2401_1

Watson, J. M. (2009). The influence of variation and expectation on the developing awareness of distribution. *Statistics Education Research Journal, 8*(1), 32–61.

Watson, J. M., & Kelly, B. (2007). Assessment of students' understanding of variation. *Teaching Statistics, 29*(3), 80–88.

Wild, C. J. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10–25.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–261.

APPENDICES

Appendix A

Recruitment Script

Dear STAT 2000 students:

I am a graduate student in the Department of Mathematics and Science Education working under the direction of Dr. Jennifer J. Kaplan in the Department of Statistics at The University of Georgia. I invite you to participate in a research study entitled "Undergraduate Students' Informal Notions of Variability." The purpose of this study is to investigate undergraduate students' ways of reasoning about the statistical concept of variability. The study will focus on ways that undergraduate students address variability in various distributions and datasets with different characteristics.

We obtained your contact information from your STAT 2000 course professor and/or computer lab teaching assistant.

In order to be eligible for this research, you must both be

1) Currently enrolled in STAT 2000 at the University of Georgia and
2) 18 years or older

Your participation will involve being interviewed in a series of interviews during the semester you are enrolled in STAT2000. You will be interviewed at most three times during the semester. Each interview will last no more than one hour so your total time commitment will be less than 3 hours. After preliminary analysis of the first interviews, candidates for the second interview will be selected. The participants for the third interview will be selected from the participants who took first and second interviews. If you are chosen for a second interview and/or third interview, the researcher will contact you to arrange the next interview.

In the interview, you will be asked to work on statistics tasks and think aloud as you solve them. The tasks will be similar to those used in class or introductory statistics textbooks. These will include tasks about the concept of variability and other ideas, concepts, and representations related to variability. The purpose of the interviews is to get more detailed information about your thinking about statistical concepts, not to grade you.

Interviews will take place outside of scheduled class meetings in an empty conference room or classroom at a time that is mutually convenient. The interviews will be video-recorded, but the camera will be focused on paper on which you will be working not on you or your face. In order to be used for the data analysis, the video recordings will be transcribed later. We plan to keep video recordings for five years, and any written work and transcripts indefinitely. These data will be stored in researcher's password-protected personal computer and on a portable hard drive.

This research involves linking your interview data with your anonymous responses to homework and lab questions from WebAssign, which we are collecting throughout the semester in the STAT 2000 course. Giving consent here includes allowing us to match your responses to online homework and lab questions of STAT 2000 to your interview data.

For the first one-hour interview, you will receive an incentive of 15 dollars at the end of the interview. If you are chosen for the second and third interviews, you will be paid 20 dollars at the end of the second one-hour interview, and 25 dollars at the end of the third interview you will have participated. If you choose to withdraw consent later, you will not lose your incentive (i.e. the payment you have already received).

The findings from this project may provide information on how to improve instruction of STAT 2000 and similar courses at other institutions. The main benefit of the interviews may be the opportunity to think more about the course material. There are no known risks associated with this research. There are some minimal discomforts associated with this research. The discomforts include the stress due to the presence of a video camera in the room or because you may be asked to explain your thinking when you are not sure whether what you are saying is statistically correct. If you feel uncomfortable, you are free to skip questions or tasks, or discontinue the interview at any time without explanation.

If you have any questions about this research project, please feel free to call me at (706) 254-6282 or send e-mail to oguzkkl@uga.edu. After I have graduated (May 2017), please direct any questions or concerns to the Principal Investigator of the study, Dr. Jennifer J. Kaplan (jkaplan@uga.edu).

Thank you for your consideration!

Sincerely,

Oguz Koklu

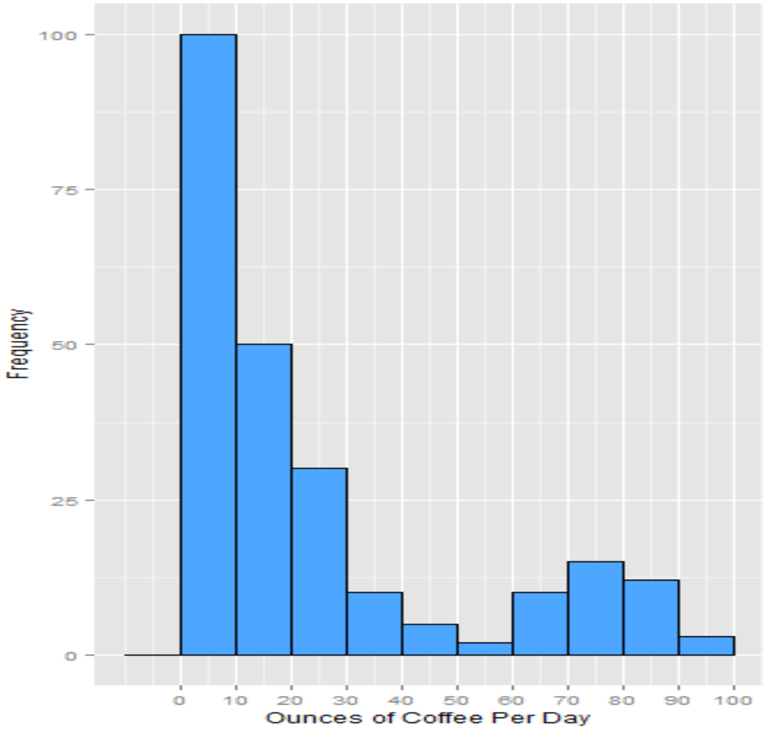
Appendix B

Homework Questions

* indicates that the question was created by the Statistics Education Research Team at UGA.

*1.

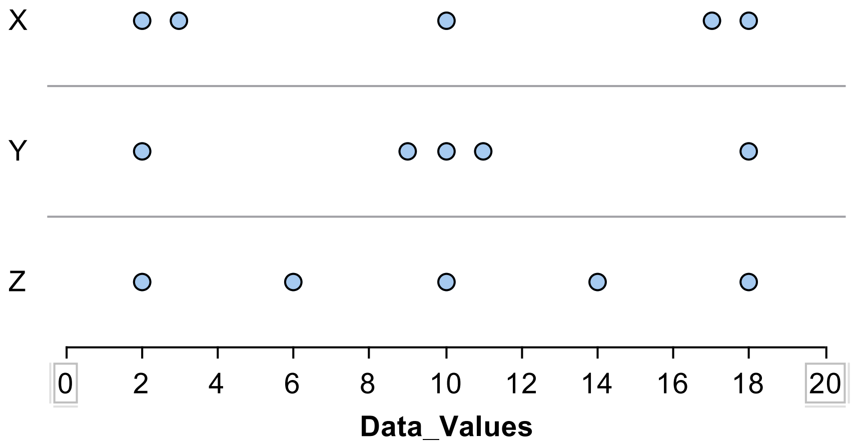
The histogram below shows the distribution of the number of ounces of coffee a random sample of 237 college students drank the previous day.



Describe the distribution of the number of ounces of coffee college students drink as shown in the histogram.

Rationale for Inclusion of the Question: This question is a typical statistical task that is commonly found in introductory statistics textbooks. For this type of questions students are supposed to describe the distribution by addressing the shape, center, and variability of the given distribution in context. I included this question in order to have an overall idea on students' reasoning about variability.

2.



Version 1

a. Which of the datasets depicted in the graph above has the least variability?

a. X

b. Y

c. Z

d. Cannot be determined.

b. Explain why you made the choice you did in part **a.**

Version 2

a. The lab assistant asked students in her morning (X), noon (Y) and afternoon (Z) open lab section hours how many times they went out to eat during the last semester and displayed the responses in the graph above. Which section has the least variable responses?

a. X

b. Y

c. Z

d. Cannot be determined.

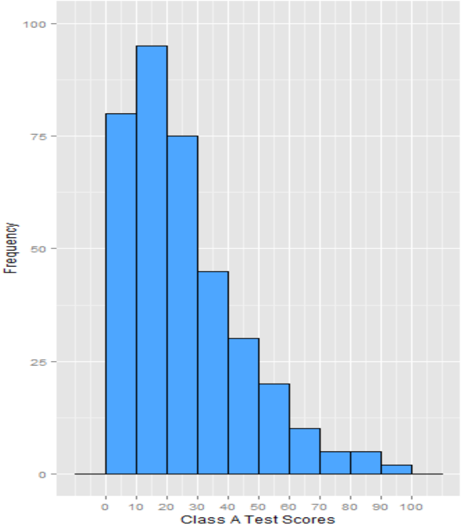
b. Explain why you made the choice you did in part **a.**

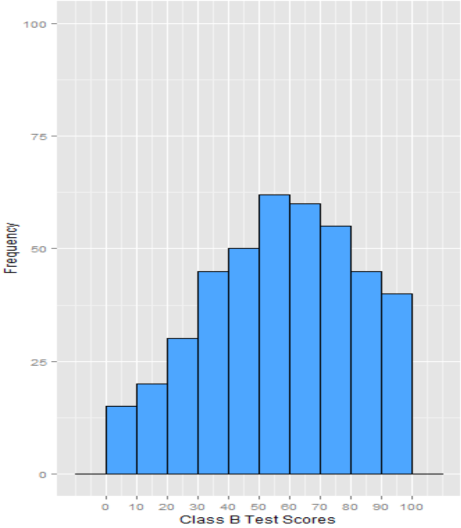
Rationale for Inclusion of the Question: This question addressed research question 1.a because the three datasets to be compared have equal ranges. In addition, it also addresses the second research question because half of the student population received the first version with no context, and the other half received the second version with context.

*3.

The histogram on the left shows the distribution of Class A's test scores for a mathematics test.

The histogram on the right shows the distribution of Class B's test scores on the same test.



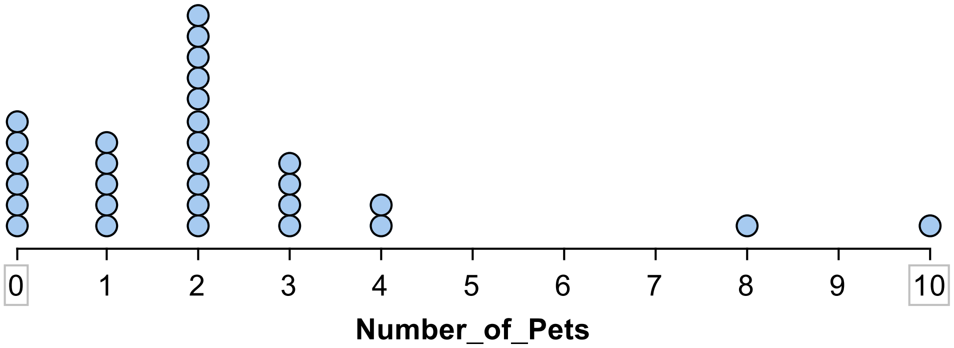


Compare the variability of the distributions of test scores for Class A (Left Histogram) and Class B (Right Histogram).

Rationale for Inclusion of the Question: This question targets research question 1.a and 1.b because the datasets to be compared have equal ranges; thus I expected students not to use *range* when reasoning about variability. In addition, it is less likely that students claim the distribution to have extreme values; hence, students' answers to the question might shed light on research question 1.b.

4.

The dot plot below shows the distribution of the number of pets owned by each of 30 students in a class.



a. Describe the distribution of the number of pets owned by the students in the class.

b. It was noticed that two of the observations, 8 and 10 as the number of previously owned pets, occurred due to an error while recording the data. These students actually had 1 and 2 pets. Given this new information, describe the new distribution of the number of pets owned by the students in the class.

Rationale for Inclusion of the Question: This question was adapted from LOCUS (Jacobbe, Case, Whitaker, & Foti, 2014), and it addresses research question 1.b. due to the extreme values 8 and 10 in the distribution. Having the question in two parts was assumed to help investigate whether the presence of extreme values leads students to solely focus on those extreme values in reasoning about variability.

5.

Version 1

a. Two datasets are given as follows:

(1st dataset) 3 3 3 40 40 40 77 77 77

(2nd dataset) 10 11 12 13 40 40 70 75 84

Without calculating, determine which dataset has more variability. Explain the reason.

Version 2

b. The researcher took a random sample of 18 law schools and randomly split the group in half. The numbers below show the percentage of graduates from each school that started to work as a lawyer upon graduation.

(1st dataset) 3 3 3 40 40 40 77 77 77

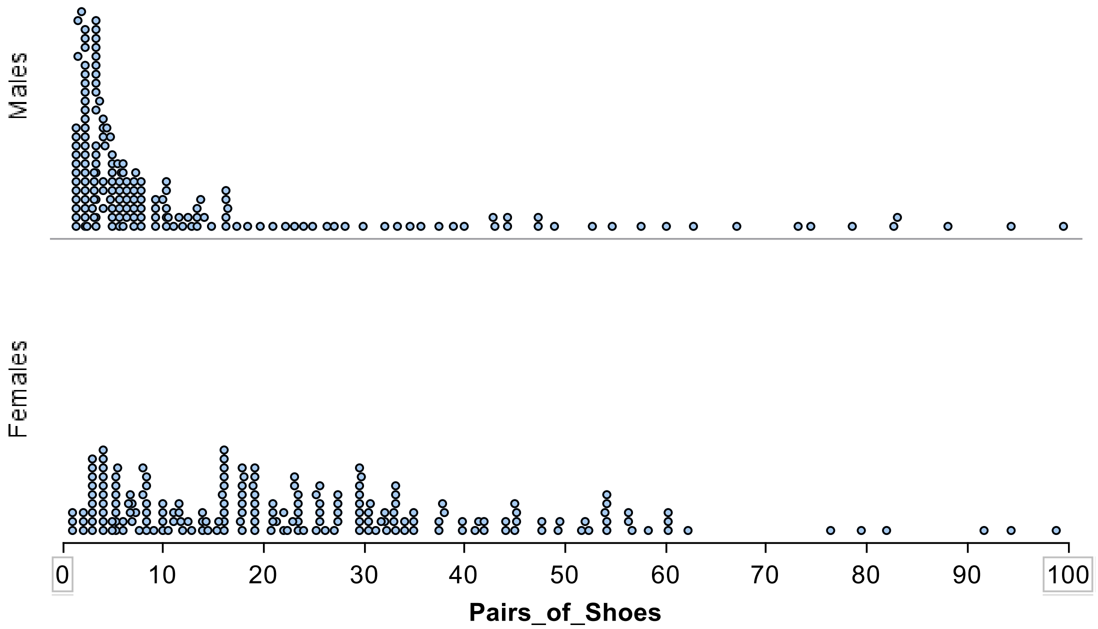
(2nd dataset) 10 11 12 13 40 40 70 75 84

Without calculating, determine which dataset has more variability. Explain the reason.

Rationale for Inclusion of the Question: This question was adapted from a question in Lann and Falk (2003). The first dataset had three different values whereas the fourth dataset has eight different data values. In addition, the datasets to be compared have equal ranges; thus, students are expected not to use range to reason about variability. In addition, it addresses the second research question because half of the student population received the first version, and the other half received the second version. Context is excluded in the first version, and included in the second version.

6.

The dot plots below show the number of pairs of shoes owned by females and males who took a survey.



a. Which group has more variability?

a. The distribution for Females is more variable.

b. The distribution for Males is more variable.

c. The variability is the same for both groups.

d. A comparison about the variability for the two groups cannot be made from the dot plots.

b. Explain your choice in the part above.

Rationale for Inclusion of the Question: This question, which was adapted from Gould and Ryan (2014, p. 64), addresses research question 1.a and 1.b. The distributions to be compared were intentionally ill constructed. Both distributions have equal ranges and it is difficult to predict which distribution has a larger standard deviation. As a result, students might employ their preexisting informal reasoning. They, for example, might claim that the distribution for females is less variable because there are only six very large values, but the great majority is spread between 0 and 60 whereas there are more values spread between 20 and 100 in the distribution for males.

*7.

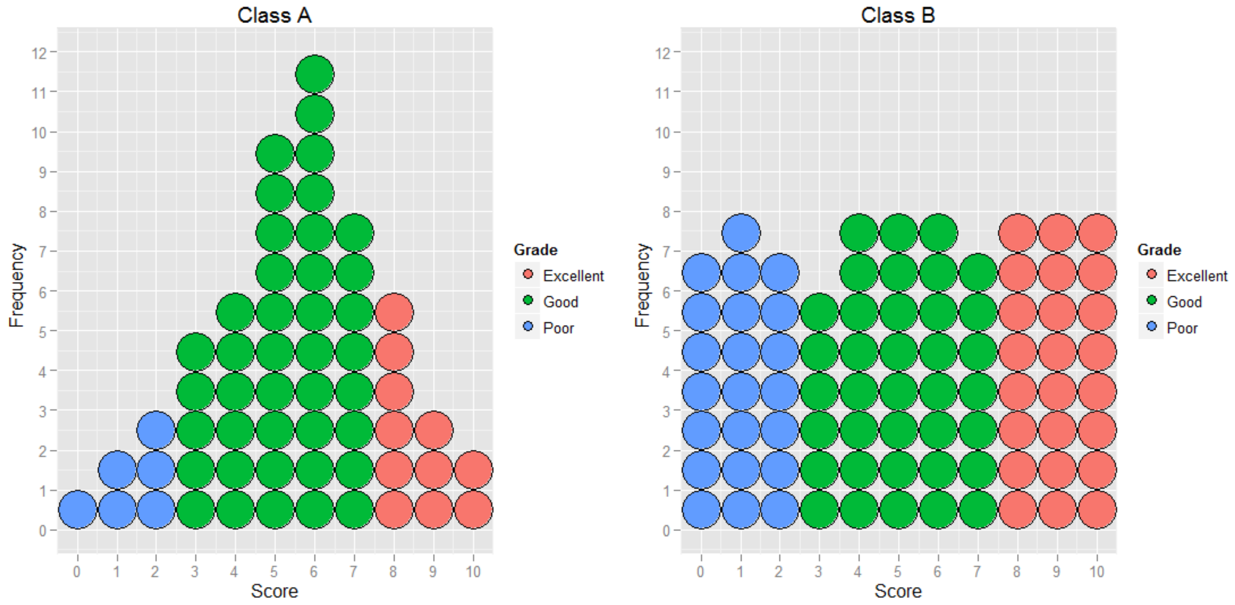
The dot plots (or histograms, depending on which version the student received) below show the distribution of scores on a 10-item test for two classes. Note that scores on each test have been classified as Excellent, Good, or Poor.

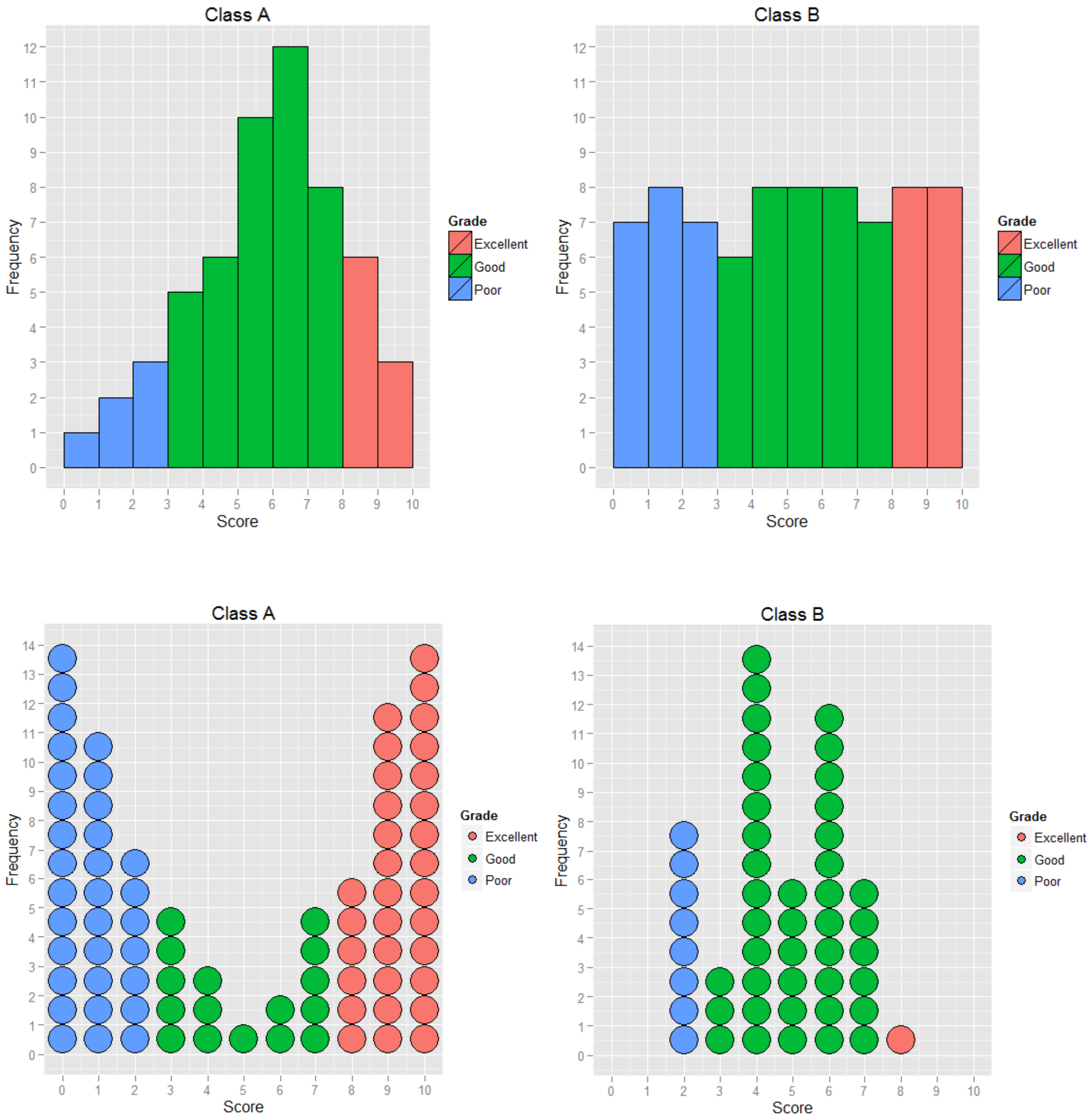
a. For which class, A or B, are the scores more variable (i.e. have the higher standard deviation)?

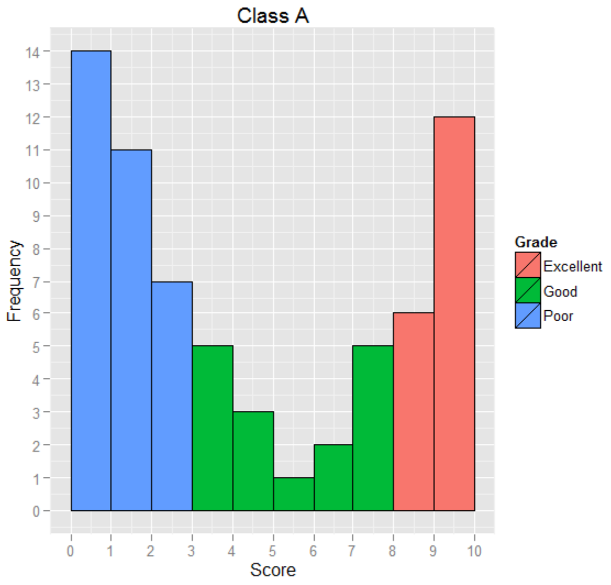
    a. Class A has more variable scores

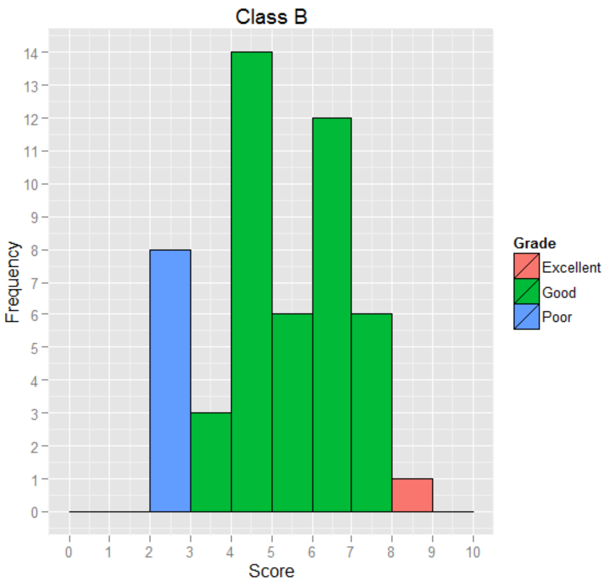
    b. Class B has more variable scores.

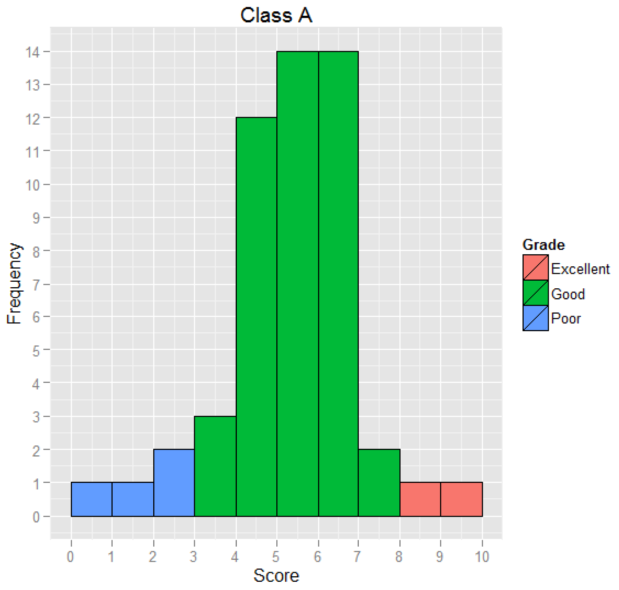
b. Explain how you know from the graphs that the scores in the class you chose are more variable.

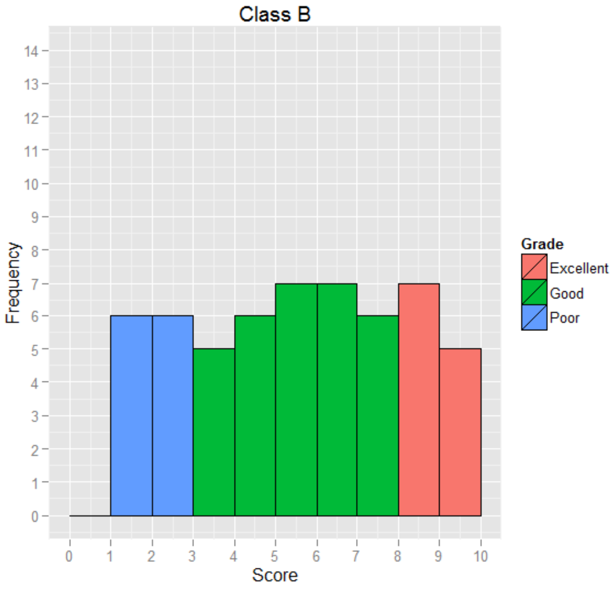


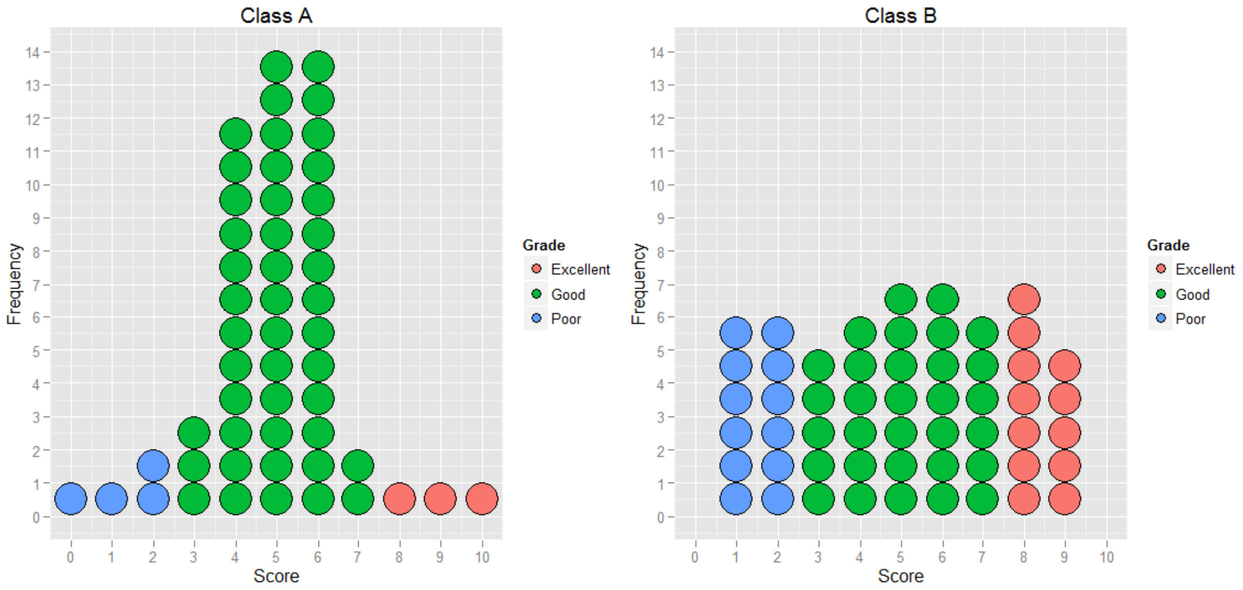

Class A
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
Score
0
1
2
3
4
5
6
7
8
9
10
Grade
Excellent
Good
Poor
Class B
Frequency
Score
Grade
Excellent
Good
Poor
Class A
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Score
Grade
Excellent
Good
Poor
Class B
Frequency
Score
Grade
Excellent
Good
Poor


Class A
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Score
0
1
2
3
4
5
6
7
8
9
10
Grade
Excellent
Good
Poor


Class B
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Score
0
1
2
3
4
5
6
7
8
9
10
Grade
Excellent
Good
Poor


Class A
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Score
0
1
2
3
4
5
6
7
8
9
10
Grade
Excellent
Good
Poor

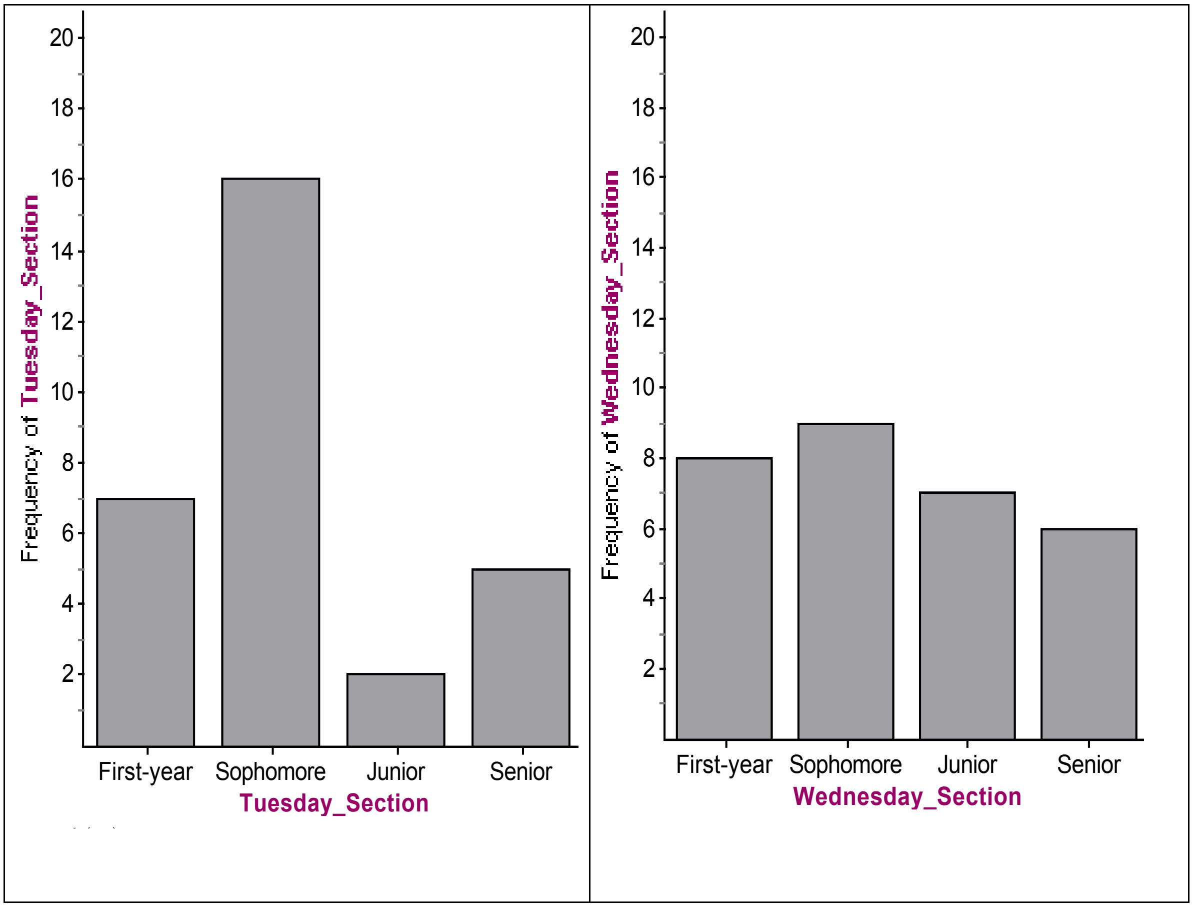

Class B
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Score
0
1
2
3
4
5
6
7
8
9
10
Grade
Excellent
Good
Poor


Class A
Class B
Frequency
Score
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Grade
Excellent
Good
Poor

Rationale for Inclusion of the Question: The stem is repeated for each pair of graphs and a student gets only a pair (version 1–6), selected randomly. This question is expected to addresses research questions 1.a, 1.b., and 1.c.

8.

The bar graphs below show the frequency distribution of class year in two sections of a statistics lab.



a. Which section has more variability in terms of class year?

a. Tuesday section

b. Wednesday section

c. The variability is the same for both sections.

d. Variability cannot be determined.

b. Explain why you made the choice you did in part **a**.

Rationale for Inclusion of the Question: The question was adapted from Gould and Ryan's (2014) introductory textbook, and it targets students' reasoning about variability for categorical data.

Appendix C

Background Questionnaire, Open-Ended Questions, Tasks

1. Thank you for volunteering to be interviewed.

2. Engage in some small talk to put subject at ease:

(What) Do you have other classes, homework or other stuff for today and this week?

3. The purpose of this interview study is to explore undergraduate college students' understanding of the statistical concept of variation and variability. I am conducting this interview on behalf of my dissertation study.

4. Do you have any experience of participating to a research like this before? How do you feel about it?

5. Are you ready for me to begin recording?

6. Today is the Month Dayth, Year, it is … o'clock. We are in room … Aderhold Hall. I am Oguz and my interviewee today is ….

1. Let's start with your background. Tell me about your (intended) major and year?
2. Tell me your experiences with statistics (both at school and in daily sense)
    a) Did you learn any statistics in high school? Before high school?
    b) Is this the first time you are taking an introductory statistics course (like STAT 2000)?
    c) If not, when did you take it before? Tell me your experiences with that class.

3. You have been enrolled to the STAT 2000 introductory statistics course and have been exposed to statistics since …. Can you tell me about that experience a little bit?

Possible follow-up questions:

a) Describe a typical stat class/lecture

b) What were some of the big take home messages in the statistics class?

c) What do you think, the professor and the TA wanted you to achieve in this class?

4. What are your thoughts about statistics as a discipline in general?

Thanks for sharing all these. Now let's focus our attention to my research study.

Open-ended Questions

1. Tell me what comes to your mind when you hear the word variability

2. How do you explain the statistical term variability?

3. Give a list of words that you think as related to variability

4. Give an example of something that "varies."

5. Give an example of something that helps you explain "variability."

6. How is the word variability used in your statistics class? Give me some examples.

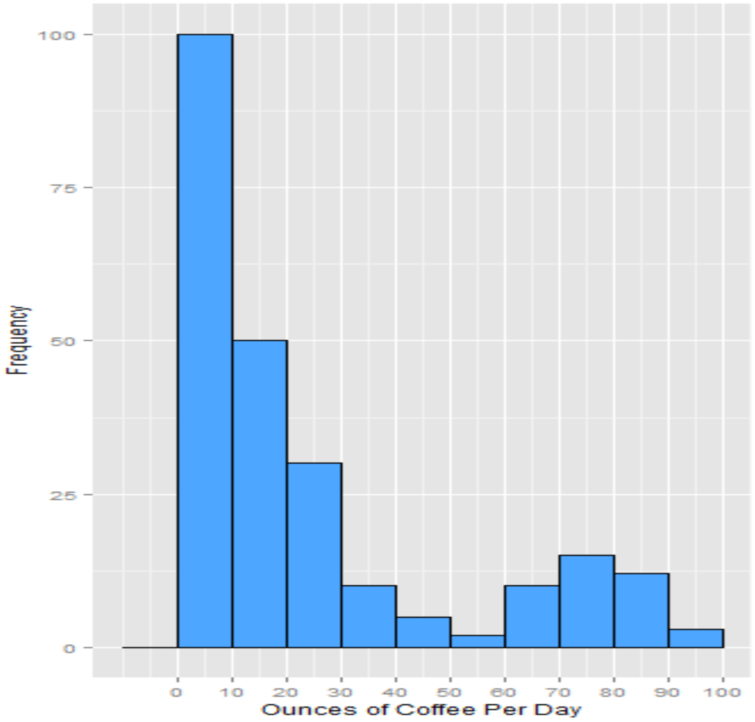
Now, I am going to ask you statistics questions. The point of the interview is for me to understand your thought processes as you work through these questions. I'd like you to think aloud as you work on and answer the questions—just tell me everything you are thinking about as you start reading and working on the questions. Don't worry whether you are right or wrong. You are welcome to use whatever tools you need to work these out (paper, calculator, StatCrunch etc.), but do your best to verbalize as much of your reasoning out loud. I may ask you follow-up questions as you answer the questions. This does not mean that you tell something right or wrong. The aim in asking these follow-up questions is to get a better sense of your reasoning.

Interview 1 Protocol

* indicates that the task was assigned as a question in the online homework assignments.

Task 1

*1. The histogram below shows the distribution of the number of ounces of coffee a random sample of 237 college students drank the previous day.



Describe the distribution of the number of ounces of coffee college students drink as shown in the histogram.

Aim: I do not aim to address the research questions specifically with this task. However, interview participants' responses to the question may provide insight on their reasoning about variability of a univariate distribution as depicted on a histogram. The follow-up question reveals the ways interview participants interpret the information vertical and horizontal axis could provide about variability.

Note: When asked to describe histogram, students are supposed to address shape, center, and variability in context. In this histogram, because the distribution is right skewed and far from being approximately normal, I do not expect many students to mention clustering around the center and notions that indicate the idea behind standard deviation. I expect to hear notions such as:

- Amount of coffee college students drank is between zero (none) and 100 ounces,
- A hundred students out of total (237) drank coffee between none to 10 ounces,
- Although the amount of coffee drank by many students is very small, there were students who drank a lot, and,
- The distribution is not normally shaped and the bins are in different length, so variability is large.

Possible follow-up questions:

1. Tell me how you used the histogram to answer the question?

2. **a.** i. What does the horizontal axis (*x*-axis) of the histogram tell?

ii. Assume that the label and the numbers in the *x*-axis are missing. Would this change your answer? If yes, how?

b. i. What does the vertical axis (*y*-axis) of the histogram tell?

ii. Assume that the label and the numbers in the *y*-axis are missing. Would this change your answer? If yes, how?

c. How is each axis related to the variability of the distribution?

3. How would variability be different if the first bin did not exist?

[If an interviewee does not address any of these]

4. What can you say about the **shape** of the distribution? What does the **shape** tell us?

5. What can you say about the **center** of the distribution?

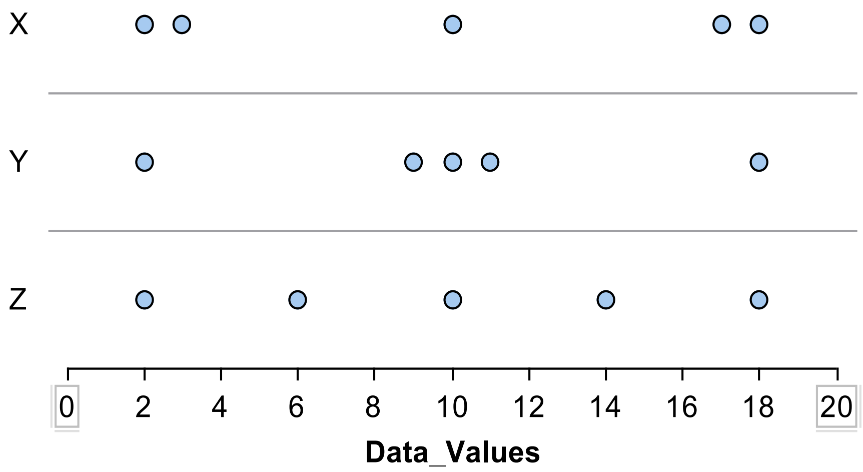
6. What can you say about the **variability** of the distribution?

a. Based on what you said about variability of the distribution, could you tell what variability means?

Task 2

*1.a. Which of the datasets, X, Y, or Z, as depicted in the graph below, has the least variability?

Explain.



Aim: The objective in asking this task is to investigate how students reason about variability if the datasets given have (approximately) equal ranges. This question specifically addresses the first research question (1.a) because the three datasets to be compared have equal ranges. In addition, it addresses the second research question because the first version of the task does not contain a context.

Note: This question was included in the homework assignments in two different ways. Half of the students took this question without a context and the second half took the question with a context. In the interviews, students are given the version without the context first. Depending on an interviewee's response, he or she is asked to come up with a context or I provide the context.

It should be noted that interviewees do not need to consider vertical axis of the dot plots because each observation in each X, Y, and Z repeats at most once (i.e., there is no frequency axis). If interviewees use the idea behind standard deviation (i.e. taking central clustering into account), then the distributions from least to most variable are Y, Z, and X in order.

Possible follow-up questions:

1. Tell me what each dot represents.

2. Tell me how you used the (given) graph to answer the question.

[If the interviewee does not explicitly tell that these are dot plots stacked together, then ask the following version of the question.]

Tell me how you used these three dot plots to answer the question.

3. Each distribution has five observations, has a mean value of 10, and range of 16(18-2=16). Considering this information, could you tell which distribution is the least variable (from the mean)?

4. a. i. What does the horizontal axis (*x*-axis) of the dot plots tell?

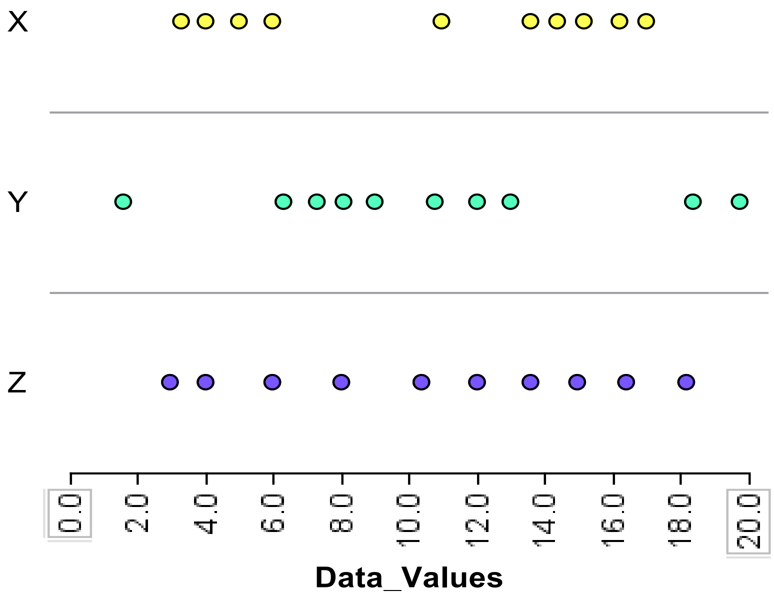
ii. Assume that the numbers in the *x*-axis are missing. Would this change your answer? If yes, how?

5. Could you come up with a meaningful context for the distributions shown in the dot plots*?*

Note: The most important criterion about the context will be to make sure that the variable suggested by the interview participant is a univariate quantitative variable. If the participant cannot approach the question or provide a context, then I present the version with the context and ask the follow-up questions above.

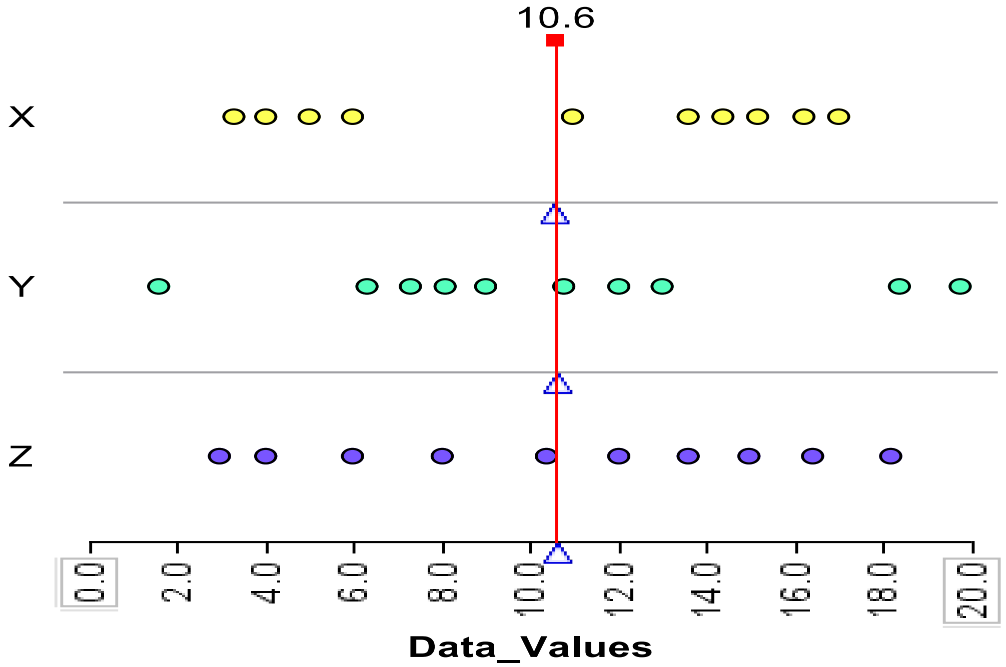
1.b. The computer-lab teaching assistant asked students in her morning (X), noon (Y) and afternoon (Z) open lab sessions how many times they went out to eat during last semester, and she displayed the responses in the dot plots above. Which section has the least variable responses? Explain.

2. Rank X, Y, and Z from least the most variable.



Note: The distributions from the least to most variable in terms of range are X, Z, and Y. However, the spread from the center is also the largest for X, and smallest for Y. As a result, using range to reason about variability should suggest a completely different conclusion as compared to using the idea of clustering around the center.

3. Distributions X, Y, and Z share the same mean value of 10.6. Y has the largest range and X has the smallest range. Rank X, Y, and Z from least the most variable.



Possible follow-up questions:

1. How do you relate the location of the mean and the variability of a distribution?

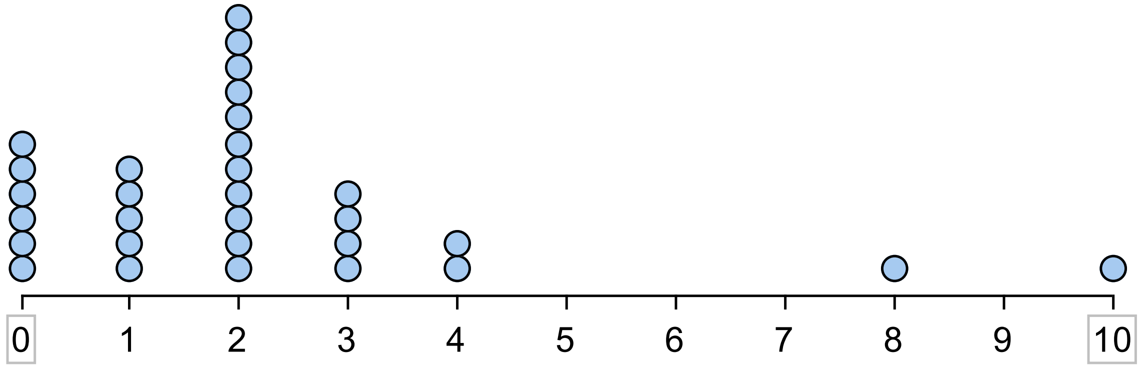
2. Which set of data varies least from the mean of the data values?

3. Which set of data has the smallest standard deviation? Explain.

4. Which set of data has the highest standard deviation? Explain.

Task 3

1.a. The dot plot below shows the distribution (of 30 observations) of a variable.



a. Describe the distribution of the variable as shown in the dot plot. Be sure to mention the variability of the distribution.

b. It was noticed that two of the observations, 8 and 10, occurred due to an error while recording the data. The values for those observations were actually 1 and 2. Given this new information, describe the new distribution of the variable. Be sure to mention the variability of the distribution.

c. Compare the variability of the distribution in the original (part a) and corrected case (part b)?

Aim: The aim of this task is to investigate research question 1.b. An interviewee's responses to part a. and part b. are supposed to show whether his or her reasoning about variability is considerably influenced by extreme values in a dataset.

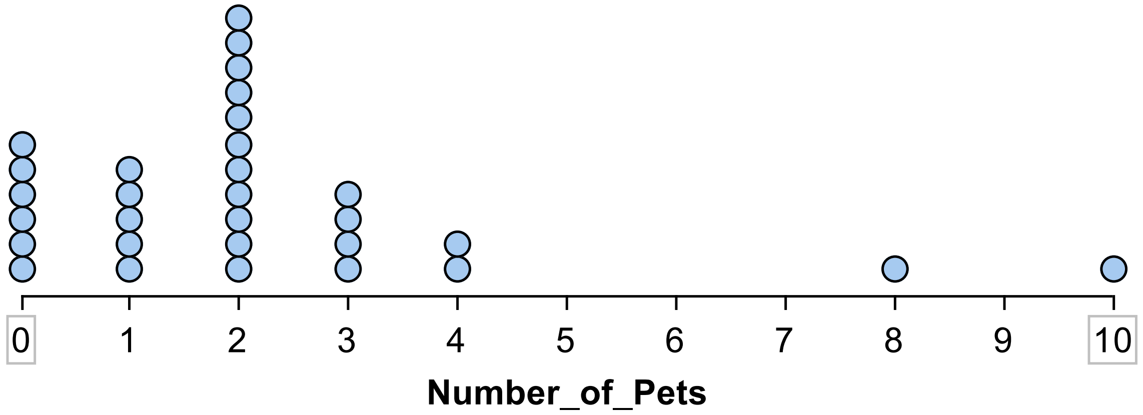
Note: Interview participants are supposed to compare the variability across the distributions and explain the reason why variability is relatively less in the later case. Claims such as "since there is no extreme value in the second case" still need further investigation. The participants are supposed to express clearly that extreme values in a dataset indicates a larger variability because they are far from a center, or they are against the clustering around center.

Possible follow-up questions:

1. Assuming the original case is correct, how would you increase the variability by adding more observations to the distribution?
2. Assuming the original case is correct, how would you decrease the variability by adding more observations to the distribution?
3. How do you relate the shape and mean of the distribution to its variability?
4. Why do you think the distribution is less variable in the corrected case?
5. Could you come-up with a meaningful context for the distributions?

Note: The task below is the context I present if a participant cannot come up with a context. Same follow-up questions will be asked.

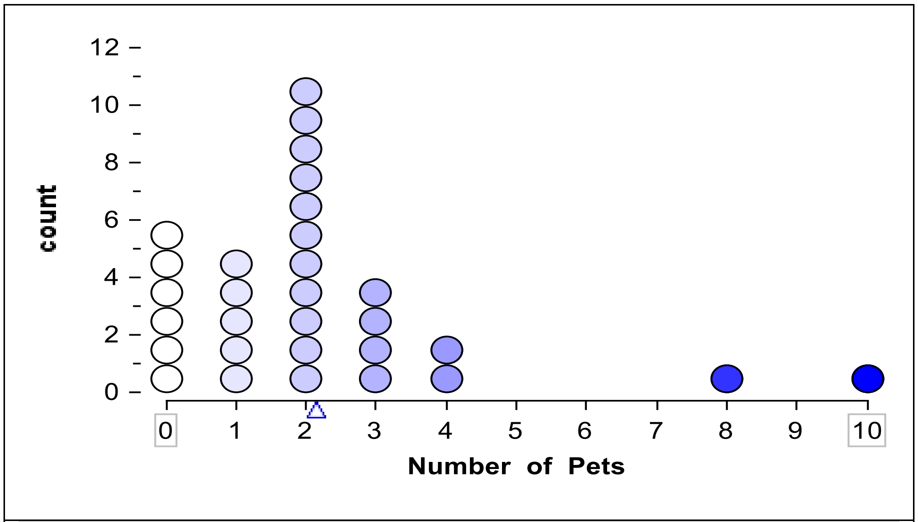
*1.b. The dot plot below shows the distribution of the number of pets owned by each of 30 students in a class.



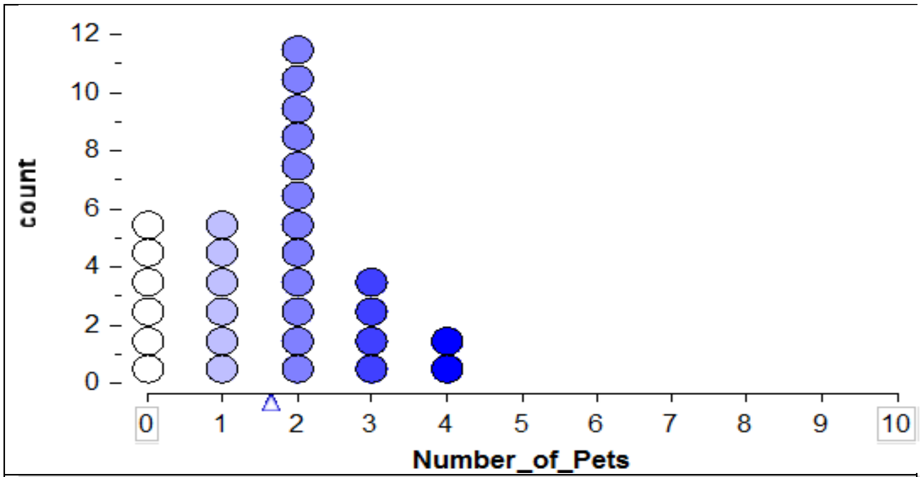
Describe the distribution of the number of pets owned by students in the class.

a. It was noticed that two of the observations, 8 and 10 as the number of previously owned pets, occurred due to an error while recording the data. These students actually had 1 and 2 pets. Given this new information, describe the new distribution of the number of pets owned by the students in the class.

1.c. Compare the variability of the distributions in the original and corrected case.



n	Mean	SD	Min.	Q1	Median	Q3	Max.	IQR
30	2.17	2.20	0	1	2	2.75	10	1.75



n	Mean	SD	Min.	Q1	Median	Q3	Max.	IQR
30	1.67	1.15	0	1	2	2	4	1

Possible follow-up questions:

1. Tell me how you use the dot plots to compare variability between the distributions?
2. **a.** i. What does the horizontal axis (*x*-axis) of the dot plot tell?

   ii. Assume that the label and the numbers in the *x*-axis are missing. Would this change your answer? If yes, how?

   b. i. What does the vertical axis (*y*-axis) of the dot plot tell?

   ii. Assume that the label and the numbers in the *y*-axis are missing. Would this change your answer? If yes, how?

   c. How is each axis related to the variability of the distribution?

[If an interviewee does not address any of these]

3. What can you say about the **shape** of the distribution? What does the **shape** tell us?
4. What can you say about the **center** of the distribution?
5. What can you say about the **variability** of the distribution?

5.a. How do you relate the location of the mean (i.e., measure of center) and the variability of a distribution?

6. In what way or ways extreme values increase the variability of a distribution?

Note: The dot plots show the frequency of each observation (*y*-axis), the location of the *mean* number of pets, and some other summary statistics. Having mean and standard deviations may be helpful for students to think about variability beyond range and existence or inexistence of extreme values.

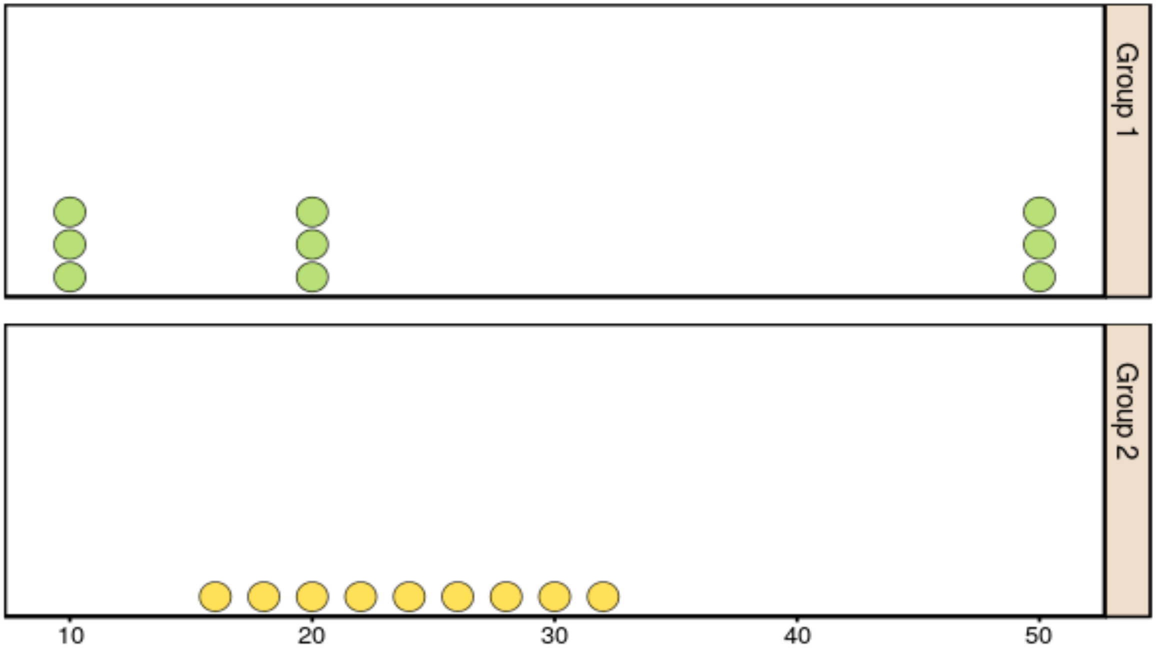
Task 4

1. Two datasets are given as follows:

(1st dataset)   10, 10, 10, 20, 20, 20, 50, 50, 50

(2nd dataset)   16, 18, 20, 22, 24, 26, 28, 30, 32

Determine whether 1st or 2nd dataset contains data that are more variable (i.e. have more variability), or if both datasets have data that are approximately equally variable. Describe why one list is more variable than the other or why they're both approximately equally variable.



Group	n	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.	IQR
1	9	27	18	10	10	20	50	50	40
2	9	24	5.5	16	20	24	28	32	8

*1.b. Two datasets are given as follows:

(1st dataset)   3 3 3 40 40 40 77 77 77

(2nd dataset)   10 11 12 13 40 40 70 75 84

Without calculating, determine which dataset has more variability. Explain.

Note: Two versions of this question are included in homework assignments. In the first version, there is no context, whereas the second version has a context. In the interviews, I will direct the version without context first. Depending on the interviewee's performance, I may provide the question in context.

Both datasets have ranges equal. In addition, the datasets do not have extreme values and I do not expect interviewees to claim any of the values as extreme. For instance, the numbers 3 in the first dataset are comparably small values but there are three of them. On the other hand, the second dataset has smaller values as 10, 11, 12, and 13, which should not appear extreme because they are sequential. The larger values in both dataset also are not expected to appear extreme due to similar reasons. As a result, I expect that interview participants do not rely on range or extreme values to reason about variability.

1.c. Without calculating, determine which dataset is more variable. Explain the reason.

(1st dataset) 3 4 10 40 40 42 42 76 77

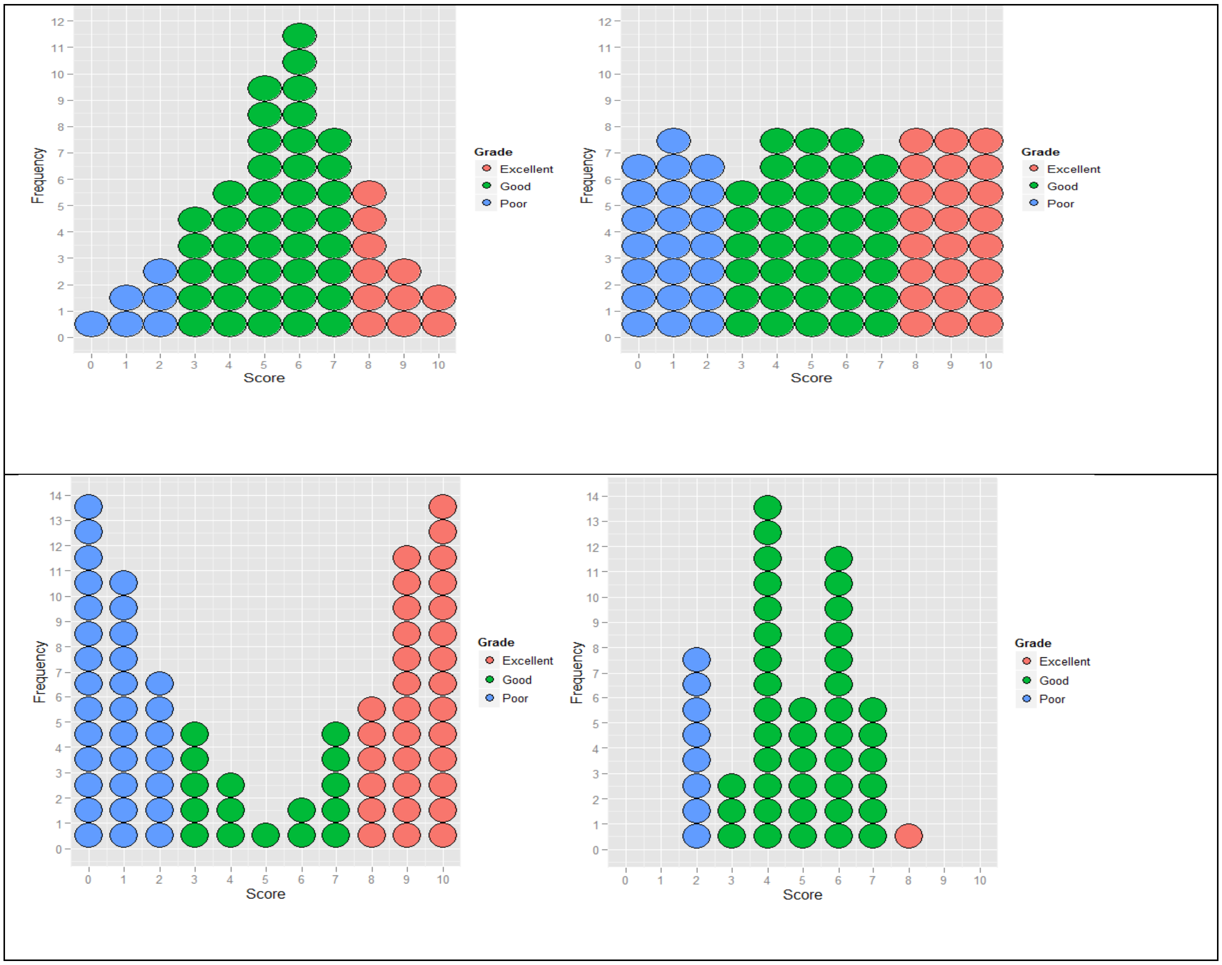
(2nd dataset) 10 11 13 13 18 70 70 75 84

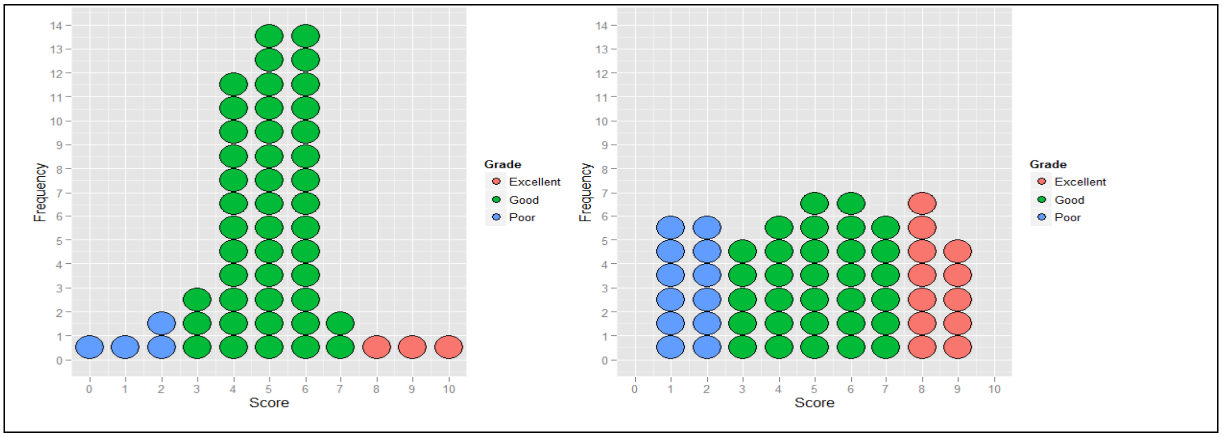
Task 5

*1. The dot plots below show the distribution of scores on a 10-item test for two classes. Note that scores on each test have been classified as Excellent, Good, or Poor.

a. Compare the distributions of scores for the classes and rank them in terms of their variability.

b. Explain how you know from the graphs that the scores in the class you chose are more variable.




Frequency
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Score
Grade
Excellent
Good
Poor
Frequency
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
Score
Grade
Excellent
Good
Poor

Note: In the homework assignment, each student had only two of the dot plots and the question explicitly asks them to think in terms of standard deviation (see Appendix A).

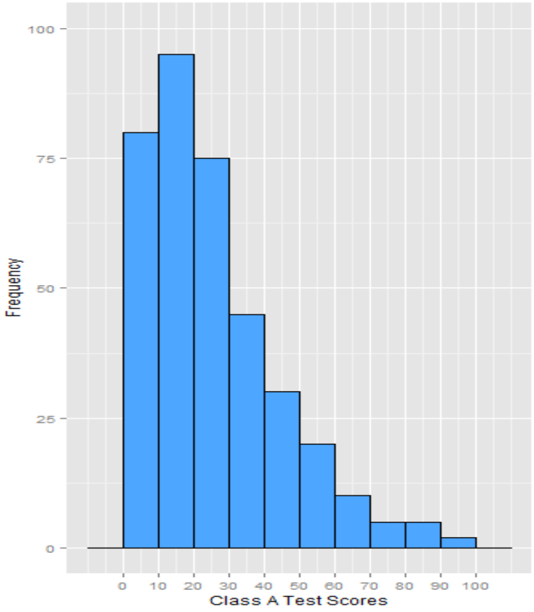
Possible follow-up questions:

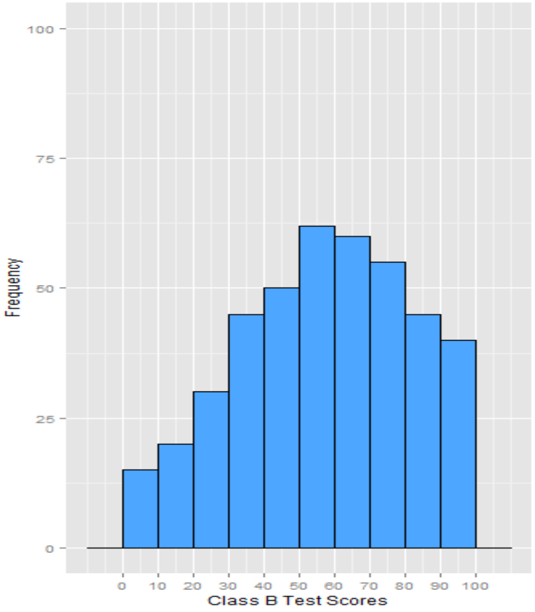
1. Which set of data varies least from the mean of scores in a class?
2. Which class has the highest standard deviation? Explain.
3. Which class has the smallest standard deviation? Explain.

Interview 2 Protocol

Task 1

*1. The histogram on the left shows the distribution of Class A's test scores for a mathematics test. The histogram on the right shows the distribution of Class B's test scores on the same test.





Compare the variability of the distributions of test scores for Class A (Left Histogram) and Class B (Right Histogram).

Possible follow-up questions:

1. What does the horizontal axis in the graphs tell you about variability?

2. What does the vertical axis in the graphs tell you about variability?

3. What does the shape of the distributions tell you about variability?

4. How do you take the mean of a distribution into account when thinking about variability?

5. Which set of data varies least from the mean of scores in a class?

Task 2

1.a

(X) 1 2 3 4 40 70 72 78 90

(Y) 7 15 30 35 40 46 47 50 96

Which of the datasets is less variable?

Group	n	Mean	Std. Dev.	Minimum	Q1	Median	Q3	Maximum	IQR
1	9	40	38	1	3	40	72	90	69
2	9	41	25	7	30	40	47	96	17

Note: Do not provide a context for the numbers in the dataset, and ask the interviewee to come up with a context.

Possible follow-up questions:

Note: If the interviewee cannot come up with a context, then provide a meaningful context.

1. X and Y stand for the two sections of the introductory Statistics course lab and the numbers stand for the distance to the home (in miles) for the students in these lab sections. Which lab section has more variability in terms of the distance to the home?

1.b. Which is less variable? X or Y? Explain.

(X) 2 3 4 10 40 70 72 78 90

(Y) 7 15 30 35 40 46 47 50 97

1.c. Which is less variable? X or Y? Explain.

(X) 2 3 4 10 16 20 40 70 72 78 90

(Y) 7 15 18 39 40 40 40 46 47 50 110

1.b.

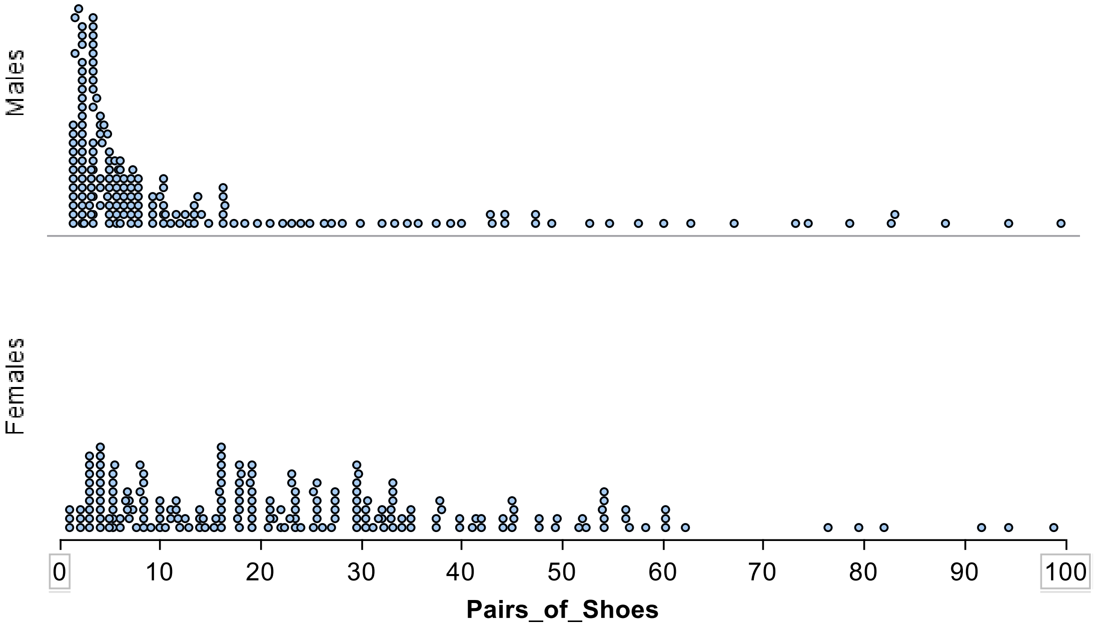
Group	n	Mean	Std. Dev.	Minimum	Q1	Median	Q3	Maximum	IQR
1	9	41	37	2	4	40	72	90	68
2	9	41	26	7	30	40	47	97	17

1.c.

Group	n	Mean	Std. Dev.	Minimum	Q1	Median	Q3	Maximum	IQR
1	11	37	34	2	7	20	71	90	64
2	11	41	27	7	28	40	46	110	18

Task 3

*1. The dot plots below show the number of pairs of shoes owned by females and males who took a survey.



a. Which group has less variability?

   a. The distribution for Females is less variable

   b. The distribution for Males is less variable

   c. The variability is the same for both groups

   d. A comparison about the variability for the two groups cannot be made from the dot plots.

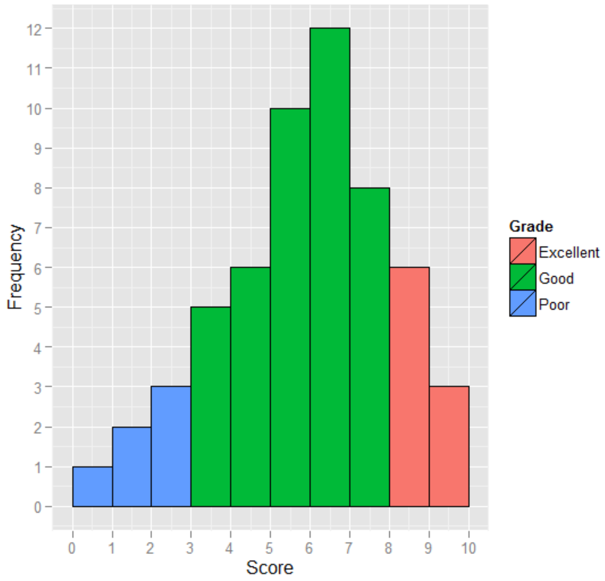
b. Explain your choice in the part above.

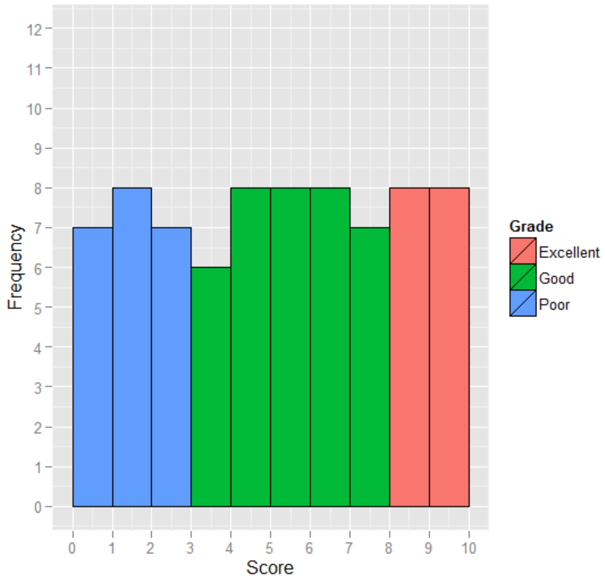
Task 4

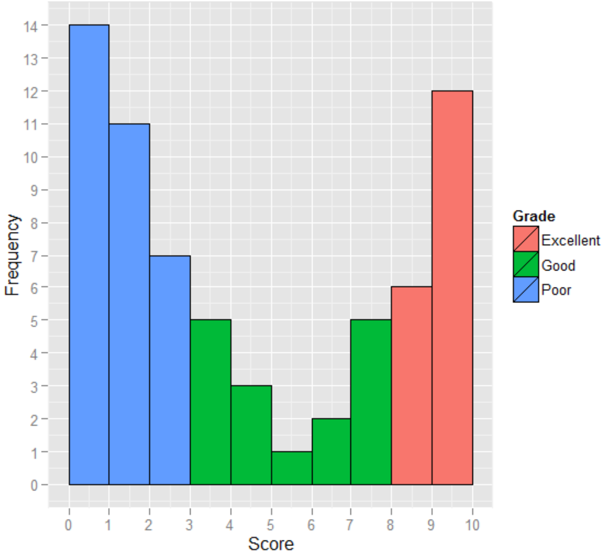
*1. The histograms below show the distribution of scores on a 10-item test for two classes. Note that scores on each test have been classified as Excellent, Good, or Poor.

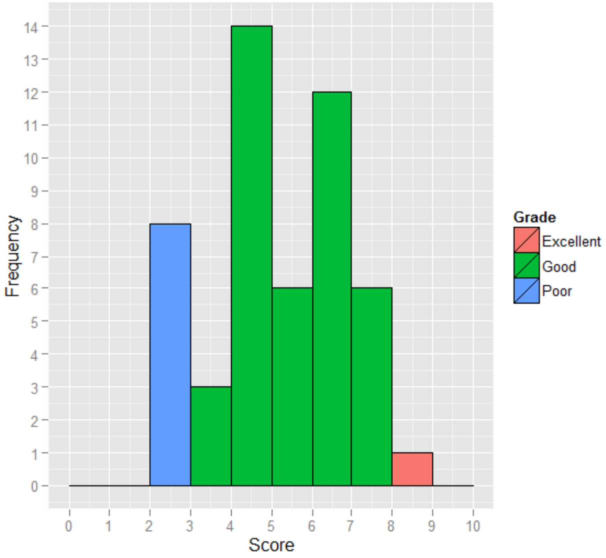
a. Compare the distributions of scores for the classes and rank them in terms of their variability.

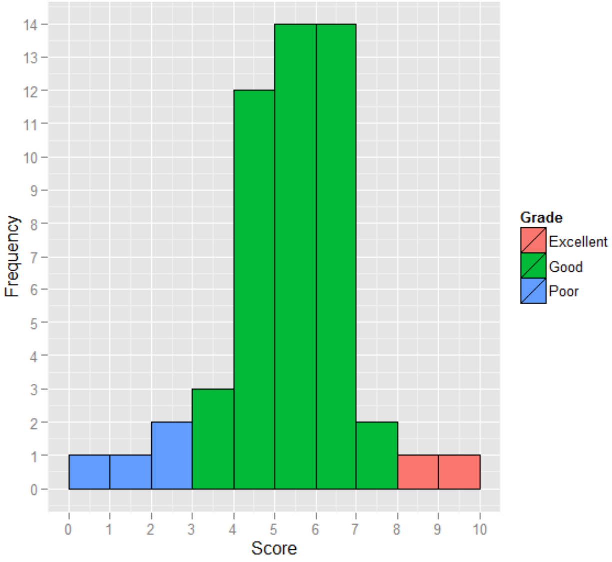
b. Explain how you know from the histograms that the scores in the class you chose are more variable.

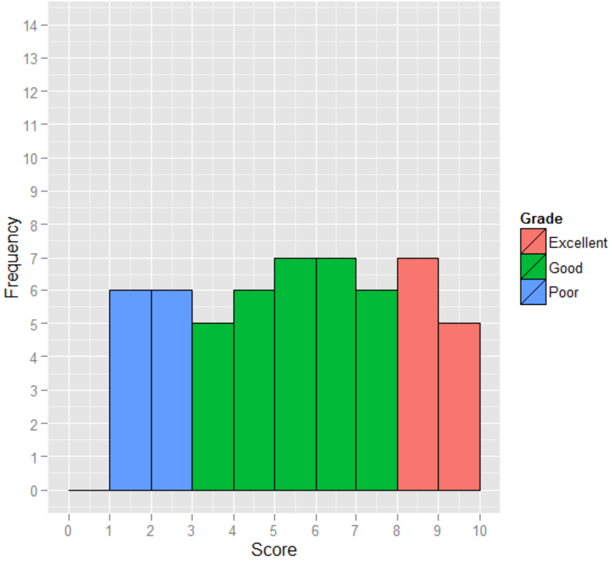





Frequency
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Score
0 1 2 3 4 5 6 7 8 9 10
Grade
🟥 Excellent
🟩 Good
🟦 Poor


Frequency
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Score
0 1 2 3 4 5 6 7 8 9 10
Grade
🟥 Excellent
🟩 Good
🟦 Poor


Frequency
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Score
0 1 2 3 4 5 6 7 8 9 10
Grade
🟥 Excellent
🟩 Good
🟦 Poor


Frequency
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Score
0 1 2 3 4 5 6 7 8 9 10
Grade
🟥 Excellent
🟩 Good
🟦 Poor

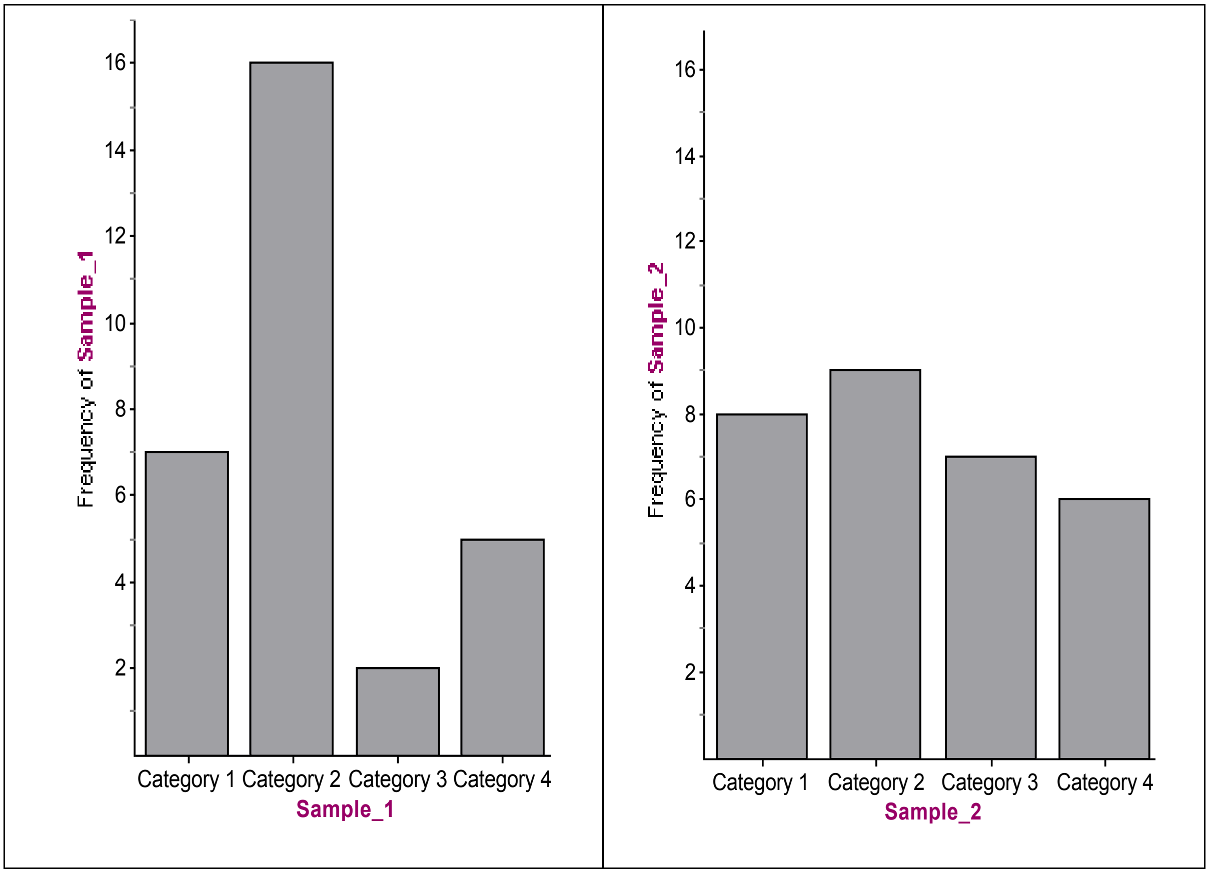
Note: In the homework assignment, each student had only two of the histograms and the question explicitly asks them to think in terms of standard deviation (see Appendix A).

Possible follow-up questions:

1. Which class seems to have the highest standard deviation? Explain.
2. Which class seems to have the smallest standard deviation? Explain.

Task 5

1.a. The bar charts below show the frequency distribution of *four categories* in two different samples.



a. Can you tell me what the bar graphs tell us?

b. Which sample has more variability in terms of the categories?

- **a.** Sample 1
- **b.** Sample 2
- **c.** The variability is the same for both samples.
- **d.** Variability cannot be determined.

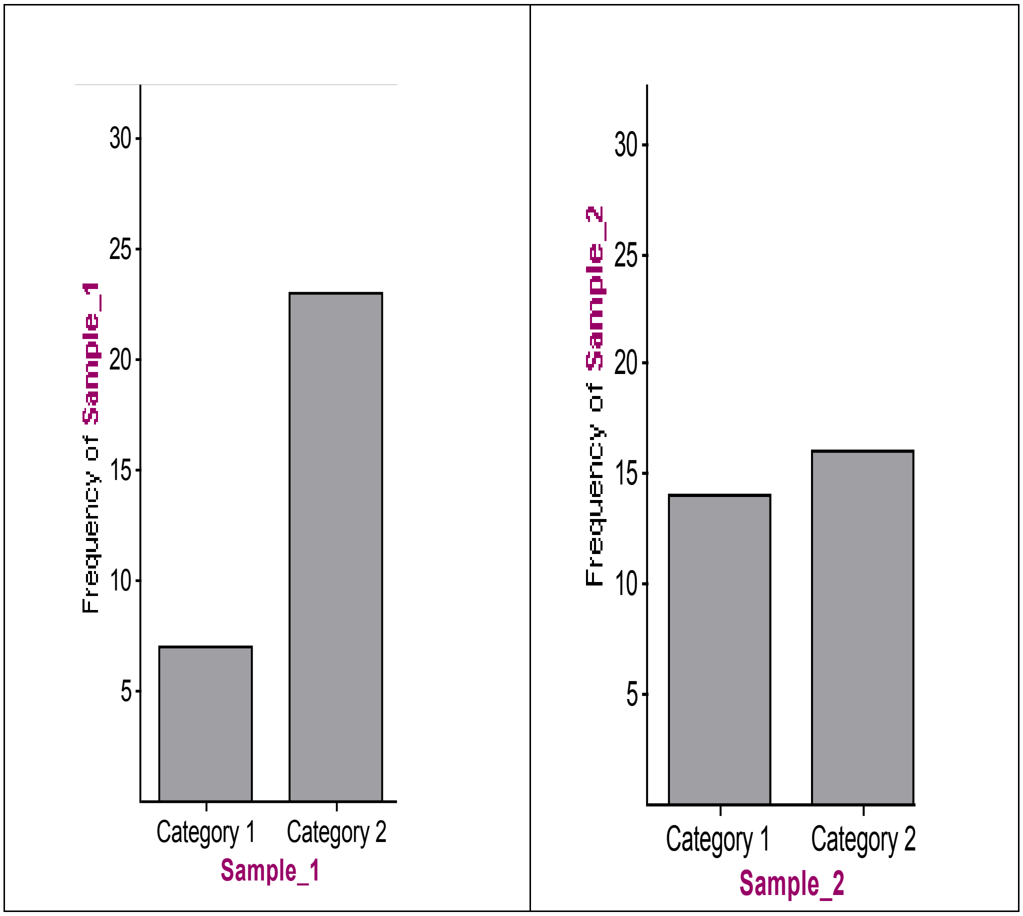
Explain why you made the choice.

Note: This task is different from the others since it focuses on reasoning about variability for categorical data. Although investigating students' reasoning about variability for categorical data was not targeted in this study, I decided to include this task because the third of the aforementioned informal notions is more appropriate to adopt when the data are categorical. The task is provided without a context. If the interviewee struggles then the same task is introduced with context.

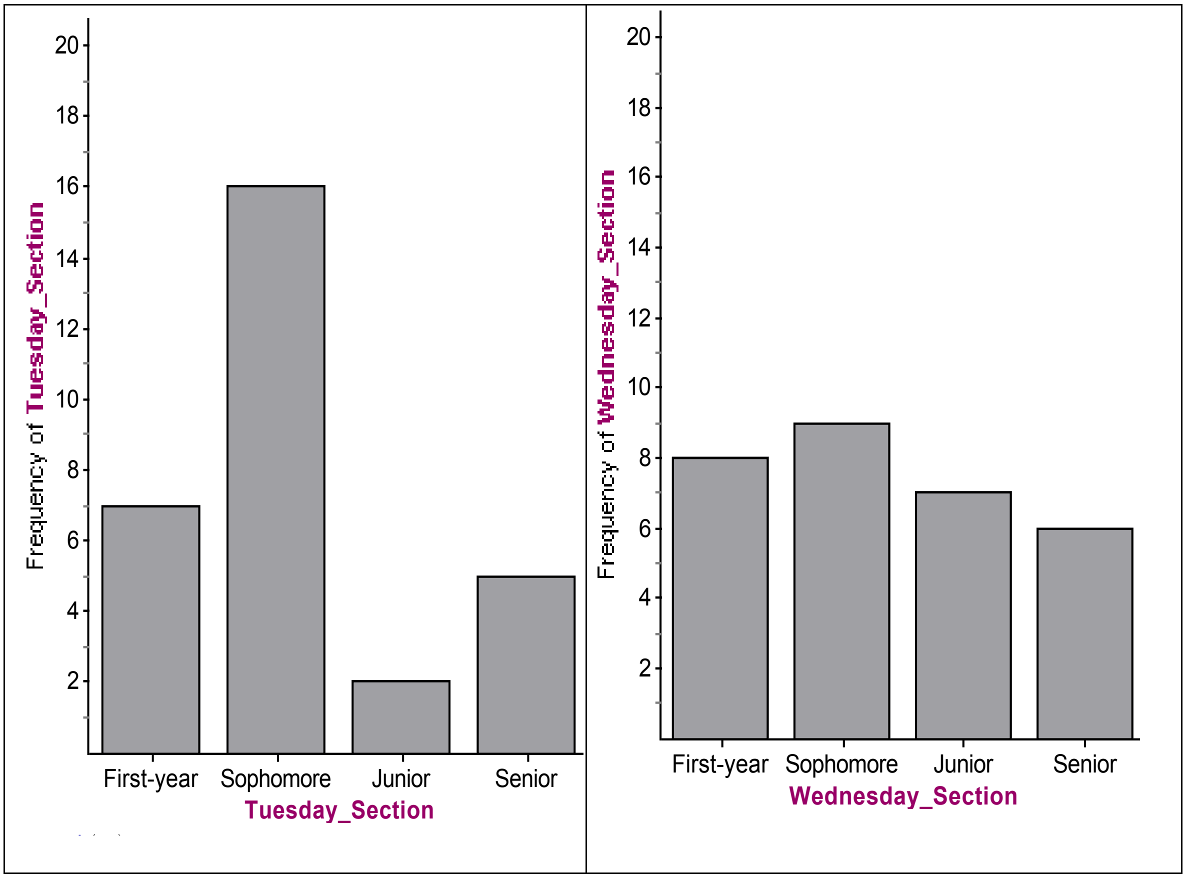
Possible follow-up questions:

1. What does the term variability refers to in this question?

Note: Having four categories may be challenging for some of the interview participants. If this is the case, I ask them to reason about variability in samples with two categories (sample size equal).


30
25
20
15
10
5
Frequency of Sample_1
Category 1
Category 2
Sample_1
30
25
20
15
10
5
Frequency of Sample_2
Category 1
Category 2
Sample_2

*2. The bar charts below show the frequency distribution of *class year* in two sections of a statistics lab.



a. Can you tell me what the bar graphs tell us?

b. Which section has more variability in terms of class year?

a. Tuesday section

b. Wednesday section

c. The variability is the same for both sections.

d. Variability cannot be determined.

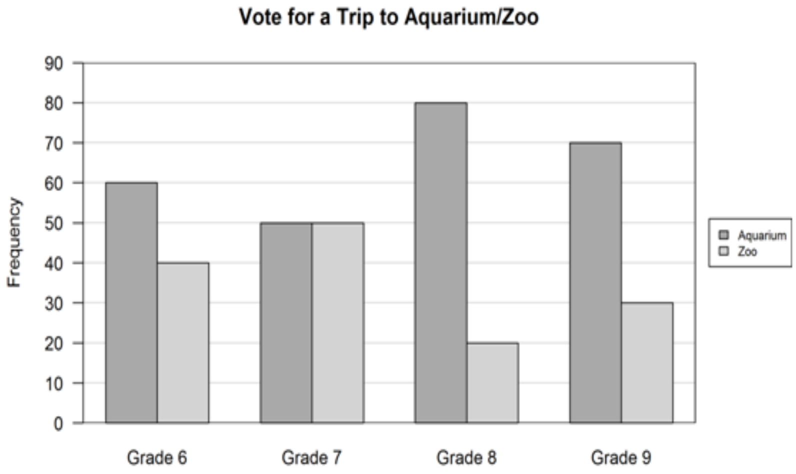
Explain why you made the choice.

3. Aquarium and Zoo

A school is planning a field trip to the aquarium or the zoo for students in grades 6–9. To determine whether the school should go to the aquarium or zoo, the school principal investigates the following statistical question:

Which field trip is most popular among students in each grade?

There are 100 students at each grade level, and every student was asked which place he or she would prefer to visit. The charts for the four grade levels are shown below.



In which grade level were the responses most variable? Explain.

Possible follow-up questions:

2. Please, tell me what the graph tell us.
3. Do you recognize any difference or similarity between the histograms and dot plots you have explored so far and this graph?
4. In which grade level were the responses less consistent? Explain.
5. What does the term variability refers to in this question.