2015

A longitudinal study of students' reasoning about variation in distributions in an introductory college statistics course

Rachel Marie Chaphalkar
The University of Montana

A LONGITUDINAL STUDY OF STUDENTS' REASONING ABOUT VARIATION IN DISTRIBUTIONS IN AN INTRODUCTORY COLLEGE STATISTICS COURSE

By

RACHEL MARIE CHAPHALKAR

Master of Science Mathematical Sciences, Michigan Technological University, Houghton, MI, 2008
Bachelor of Science Mathematics, Michigan Technological University, Houghton, MI, 2006

Dissertation
presented in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in Mathematics Education

The University of Montana
Missoula, MT

December 2014

Approved by:

Sandy Ross, Dean of The Graduate School
Graduate School

Ke Wu, Chair
Mathematical Sciences

Bharath Sriraman, Co-Chair
Mathematical Sciences

James Hirstein
Mathematical Sciences

David Patterson
Mathematical Sciences

David Erickson
Curriculum and Instruction

Chaphalkar, Rachel, Ph.D., Fall 2014 Mathematics Education

Abstract: A Longitudinal Study of Students' Reasoning About Variation in Distributions in an Introductory College Statistics Course

Chairperson: Ke Wu
Co-Chairperson: Bharath Sriraman

Current curricular documents including the Common Core State Standards (2010) and the Guidelines for Assessment and Instruction in Statistics Education (2005) have increased the need for students' understanding and reasoning about statistics at both the K-12 and college levels. In addition, an increasing number of students are taking the Advanced Placement Statistics Exam (College Board, 2011) or a college-level introductory statistics course (Scheaffer & Stasny, 2000). One of the main components for statistical thinking is consideration of variation (Wild & Pfannkuch, 1999). Previous studies have shown that students have misconceptions about variation (e.g. Reading, 2004; Torok & Watson, 1999) and students often lack the ability to give sophisticated answers (Shaughnessy, 2007). The goal of this study was to better understand how students' reasoning about variation in a distributional context changes as they progress through an introductory college-level statistics course. In order to better understand the longitudinal nature of this process during a semester-long introductory statistics course, both quantitative and qualitative data were collected at three different times (beginning, middle, and end of the course) in surveys and interviews. The Structure of Observed Learning Outcomes (SOLO) Taxonomy (Biggs & Collis, 1982) was used to understand and assess the quality of their reasoning. Qualitative data came from two sources: three interviews from each of the ten interviewees and three survey questions on each of three surveys from all participants. The interviews were transcribed and responses were sorted into appropriate locations in the SOLO Taxonomy. After coding responses to each question in each interview, themes of progress were then identified. These themes showed that students progressed through four different paths of reasoning including: improved, maintained, decreased, and inconsistent. Quantitative data showed that while students were good at reasoning about situations involving bar graphs and dot plots with regards to comparing variability in distributions, they struggled with reasoning about histograms. Overall, this study found that there was no statistically significant improvement in reasoning about variability when comparing distributions as students progressed through a college-level introductory statistics course. This lack of improvement suggested that perhaps college students needed to have direct intervention or cognitive conflict in order to make more progress in reasoning about variability when comparing distributions.

Acknowledgements

Many people have helped and made this dissertation possible from many different viewpoints, likely more than I will be able to mention here. I think that this all began back during my freshman year of college, as I found out that engineering (although it included mathematics and science) was not going to be the career for me. My parents were actually quite excited that I switched to a degree in teaching mathematics. This grew into a master's degree so that I could teach calculus right away (not that this happened exactly in that manner) and eventually turned into a dissertation to complete a doctoral degree in mathematics education. Of course, it did not happen that quickly or without challenges along the way, including the passing of my mother and a significant medical obstacle of my own.

Not only did my advisor, Ke Wu, provide significant educational support, I also must thank her for continuing to remind me to take care of myself. My co-advisor, Bharath Sriraman, has continued to challenge me to be a better student and researcher throughout my graduate school career. David Patterson helped get me interested in statistics education by suggesting and participating in an independent study course, as well as getting me to love theoretical statistics in one of the most challenging courses I have taken. Jim Hirstein has helped me to see manipulatives as an amazing tool for students (and pre-service teachers) and to find fun ways to teach elementary mathematics. David Erickson has provided a new perspective of mathematics education from the Curriculum and Instruction side, challenged me to understand theoretical ideas, and helped me to find a great teaching internship at a local high school.

Other faculty at University of Montana have also been supportive both in my education, research, and service activities, especially: Jenny McNulty, Jon Graham, Cindy Leary, Bonnie Spence, Matt Roscoe, Emily Stone, Kelly McKinnie, and Gretchen McCaffry. In addition, I have certainly benefited from friendships with other graduate students in the program, in many different areas of mathematics, and from the curriculum and instruction department. Now I am benefitting from my new colleagues at the University of Wisconsin – Whitewater, both in the Department of Mathematics and throughout campus.

Last but not least, my family and friends have been a big support, especially through some of the difficult life challenges I've experienced in the past five years; my husband, Nik, my dad, Kevin, my sister, Heather, my cousin, Bekah, and my friends, especially Peggy, Mary Beth, and Lisa. Also, some others from the local community, who have helped me to get through new challenges: Annie, Catherine, Jill, and Cindy. My mom, for her support at the beginning of this process, inspiration for education in the first place, and for showing me that you must go all out to make it through life's challenges.

Table of Contents

Page

Chapters

List of Tables

Table Page

List of Figures

Figure Page

Chapter 1

Introduction to the Problem

This chapter discusses the importance of statistics education for students from elementary school to college, as well as gives an overview of several curricular documents outlining what should be included in a students' statistical education. Two concepts, variation and distribution, which are essential to statistical thinking and this dissertation, are defined and situated within statistics. Finally, several challenges in the area of students' learning of statistics are discussed, specifically those focused on variation and distribution.

Importance of Statistics Education

The study of statistics is becoming increasingly important in both K-12 education and at the college level; this can be seen in its inclusion in curriculum documents, the presence of more statistics items on national assessments, and an increase in college course enrollment in introductory statistics courses.

Curriculum documents. Since 1989, there have been a number of curricular documents increasing the role of probability and statistics in education from elementary through high school and at the college level. The National Council of Teachers of Mathematics (NCTM) first introduced statistics as a content strand and enumerated the appropriate statistics material for K-12 students (1989, 2000). The American Statistical Association published the Guidelines for Assessment and Instruction in Statistics Education (GAISE) directed at both Pre-Kindergarten through Grade 12 and post-secondary statistics instruction (Garfield, Aliaga, Cobb, Cuff, Gould et al., 2005; Franklin, Kader, Mewborn, Moreno, Peck et al., 2007). Most recently, the Common Core

State Standards (CCSS) (2010) include probability and statistics as a content domain for 6th through 8th grade and conceptual category for high school.

Assessment. Shaughnessy and Zawojewski (1999) found an increase in the number of questions on the National Assessment of Educational Progress (NAEP) about statistics and probability, as well as an increase in teachers and students reporting teaching and learning probability and statistics on the NAEP survey. Student scores on probability and statistics items had a statistically significant increase from 1990 to 1996. Shaughnessy and Zawojewski (1999) reported that all grades taking the NAEP exam (fourth, eighth, and twelfth) had difficulties providing reasoning when interpreting graphs, and twelfth grade students had difficulty finding the probabilities of compound events, with even fewer correctly explaining their reasoning.

College statistics course enrollment. From 1990 to 2000, enrollment in elementary statistics courses increased from 87,000 to 136,000 students in mathematics departments, 30,000 to 54,000 students in statistics departments, and 54,000 to 74,000 students in two-year colleges (Scheaffer & Stasny, 2004). For the fall semester of 2010, the Conference Board of Mathematical Sciences (CBMS) estimated 231,000 students took an introductory statistics course in mathematics departments while a very similar number of students (233,000) took mainstream Calculus I. An additional 81,000 students enrolled in introductory statistics courses through statistics departments and 134,000 students were enrolled in elementary statistics courses at two-year colleges (Blair, Kirkland, & Maxwell, 2013). Totaling the students from all of these venues, nearly twice as many students took an introductory statistics course in Fall 2010 as compared to Fall 2000. In addition, an increasing number of students are receiving college credit for an introductory statistics course by taking and scoring well on the Advanced Placement (AP) exam. The AP Statistics test was first given in 1997 and since then participation has increased from 7,500 students the first year to 142,000 students in 2011 (College Board, 2011).

Curriculum Standards

Beginning with NCTM's Standards for School Mathematics (1989, 2000), several curriculum documents have advanced topics in statistics and probability for both K-12 students and introductory college-level statistics students. Recent documents describing what should be taught include the Guidelines for Assessment and Instruction in Statistics Education (GAISE) (Garfield et al., 2005; Franklin et al., 2007) at both the K-12 and college-level, and the Common Core State Standards for Mathematics (2010) for K-12.

GAISE standards for K-12. The GAISE standards recommend that "every high school graduate should be able to use sound statistical reasoning to intelligently cope with the requirements of citizenship, employment and family and be prepared for a healthy, happy, and productive life" (Franklin et al., 2007, p. 1). In addition to containing appropriate classroom activities, the standards include a framework for the recommended statistical content and process at three levels (see Table 1). Students first gain experience with the statistical process (formulating questions, collecting and analyzing data, and interpreting results) in level A, then move to level B where the process is refined, and finally to level C where students are able to make generalizations. In addition, the framework includes descriptions of the different possible natures of variability and where students may recognize this variability.

Table 1

GAISE Statistical Process Framework for K-12 students adapted from Franklin, Kader, Mewborn, Moreno, Peck et al. (2007, p. 14-15)

Process Component	Level A	Level B	Level C
I. Formulate Question	Beginning awareness of the statistics question distinction Teacher poses questions of interest Questions restricted to the classroom	Increases awareness of the statistical question distinction Students begin to pose their own questions of interest Questions are not restricted to the classroom	Students can make the statistics questions distinction Students pose their own questions of interest Questions seek generalization
II. Collect Data	Do not yet design for differences Census of classroom Simple experiment	Beginning awareness of design for differences Sample surveys; begin to use random selection Comparative experiment; begin to use random allocation	Students make design for differences Sampling designs with random selection Experimental designs with randomization
III. Analyze Data	Use particular properties of distributions in the context of a specific example Display variability within a group Compare individual to individual Compare individual to group Beginning awareness of group to group Observe association between two variables	Learn to use particular properties of distributions as tools of analysis Quantify variability within a group Compare group to group in displays Acknowledge sampling error Some quantification of association; simple models for association	Understand and use distributions in analysis as a global concept Measure variability within a group. Measure variability between groups Compare group to group using displays and measures of variability Describe and quantify sampling error Quantification of association; fitting of models for association
IV. Interpret results	Students do not look beyond the data No generalization beyond the classroom Note difference between two individuals with different conditions Observe association in displays	Students acknowledge that looking beyond the data is feasible Acknowledge that a sample may or may not be representative of the large population Note the difference between two groups with different conditions Aware of distinction between observational study and experiment Note differences in strength of association Basic interpretation for models for association Aware of the distinction between association and cause and effect	Students are able to look beyond the data in come contexts Generalize from sample to population Aware of the effect of randomization on the results of experiments Understand the difference between observational studies and experiments Interpret measures of strength of association Interpret models of association Distinguish between conclusions from association studies and experiments
Nature of Variability	Measurement variability Natural variability Inductive variability	Sampling variability	Chance variability
Focus on Variability	Variability within a group	Variability within a group and variability between groups Covariability	Variability in model fitting

Common Core State Standards. As of this writing, 43 out of 50 states have adopted the Common Core State Standards for Mathematics, with full implementation typically scheduled for the 2013-2014 or 2014-15 school year (CCSS, 2010). Although the Probability and Statistics strand does not start until Grade 6, the focus of "measurement and data" for elementary school students gives students the tools that will be used for meaningful statistics in later grades. Students in Grade 1 should be able to organize, interpret, and collect categorical data for up to three categories, and students in Grades 2 and 3 should be able to make pictographs and bar graphs (CCSS, 2010).

The middle school curriculum for probability and statistics includes many of the topics also included in a college-level introductory statistics course. In Grade 6, students should develop understanding of statistical variability, summarize distributions, and describe distributions. Grade 7 includes many topics: use random sampling to draw inferences about a population, draw informal comparative inferences about two populations, investigate chance processes, and develop, use, and evaluate probability models. In grade 8, students should investigate patterns of association in bivariate data (CCSS, 2010).

While specific skills are to be taught at certain grade levels for elementary and middle school students, there is more flexibility in the high school standards: they should be covered before graduation and no particular ordering of the curriculum is given for these standards. The standards (CCSS, 2010) for high school grades consist of the following clusters: (a) calculate expected values and use them to solve problems, (b) use probability to evaluate outcomes of decisions, (c) summarize, represent, and interpret one or two categorical and/or quantitative variables and linear models, (d) understand and evaluate random processes underlying statistical experiments, (e) make inferences and justify conclusions from sample surveys, experiments and observational studies, (f) understand independence and conditional probability and use them to interpret data, and (g) use the rules of probability to compute probabilities of compound events in a uniform probability model (p. 80).

GAISE standards for college-level introductory statistics courses. Introductory statistics courses are taught by many different departments with students ranging in background and abilities. The goals for students completing the courses range from being quantitatively literate to being able to produce statistical analyses. While there is a greater demand in industry for data analysis, software and other technologies are making calculations easier. This trend decreases the need to teach these calculations while increasing the importance of conceptual understanding of statistical tools so that they are used correctly. The Guidelines for Assessment and Instruction in Statistics Education College Report (Garfield et al., 2005) make the following recommendations for introductory college statistics courses: (a) emphasize statistical literacy and develop statistical thinking, (b) use real data, (c) stress conceptual understanding, rather than mere knowledge of procedures, (d) foster active learning in the classroom, (e) use technology for developing conceptual understanding and analyzing data, and (f) use assessments to improve and evaluate student learning (p. 4). In addition, topics in an introductory statistic course should include experimental design and conclusions that can be made from different types of studies, importance of randomness, exploratory data analysis, and the basic ideas of statistical inference. While it is important to cover these topics and some statistical techniques, “specific techniques are not as important as the knowledge that comes from going through the process of learning them" (Garfield et al., 2005, p. 11).

Course Sequencing. While there is agreement on which topics should be taught in an introductory statistics course, the order in which these topics should be taught is not agreed upon. Chance and Rossman (2001) give arguments for and against each of these options in sequencing: beginning the course with either exploratory data analysis or data collection, discussing bivariate data before or after inference for a single variable, studying either proportions first or studying proportions and means concurrently when beginning inference, and studying tests of significance or confidence intervals also when beginning inference. Chance and Rossman (2001) make it clear that they want students to have conceptual knowledge of statistics, not just computational knowledge, in each of these topics.

Statistical Variation and Variability

Throughout all of the topics in a statistics course, the common theme is variation. Wild and Pfannkuch (1999) described the importance of variation in statistics as: "variation is omnipresent; variation can have serious practical consequences; and statistics give us a means of understanding a variation-beset world" (p. 235). Essentially, statistics must be used due to the presence of variation in data. One of the main components of Wild and Pfannkuch's model of statistical thinking is consideration of variation. They described that consideration of variation includes: noticing and acknowledging variation, measuring and modeling variation for the purposes of prediction, explanation, or control, explaining and dealing with variation, and investigative strategies. Shaughnessy and Pfannkuch (2002) illustrated the beauty of variation in the context of Yellowstone's Old Faithful Geyer eruption wait times: a histogram of wait times shows a bimodal distribution, while a time series graph shows that these wait times tend to oscillate between the two modes. If one just described the distribution using a measure of center, the patterns in the data would be lost; without multiple graphs of different types, only certain patterns would be seen. This example illustrates how the choice of graphical display affects the pattern seen.

Statistical variation and variability are concepts that are encountered in an introductory statistics course. Gould (2004) defined variation as, "that which is not pattern" (p.8). Variation is also described as the noise in a data set (Konold & Pollatsek, 2002). A textbook written by Rossman and Chance (2008) defines variability as, "variability refers to the phenomenon of a variable taking on different values or categories from observational unit to observational unit" (p. 4). Spread is a common synonym for variability and is defined as "how much the data vary around the center" (DeVeaux, Velleman, & Bock, 2014, p. 52). In an introductory statistics course there are three measures of variability typically presented for quantitative data: standard deviation, range, and interquartile range. The standard deviation (denoted by *s*) is the square root of the variance:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}},$$

where x_i is each observation, $\bar{x}$ is the mean of the observations, and *n* is the sample size. Rossman and Chance (2008) give the following conceptual explanation: "The standard deviation can loosely be interpreted as the typical distance that a data value in the distribution deviates from the mean" (p. 164). The range is the difference between the maximum and minimum values in a data set. Rossman and Chance (2008) preface the definition of range with the caveat that it is not especially useful as a measure of variation. Lastly, the interquartile range (IQR) is defined as the difference between the upper and lower quartiles. The upper quartile is the value such that 75% of the data are below it, while the lower quartile is the value such that 25% of the data are below it. Conceptually, "the IQR is the range of the middle 50% of the data" (Rossman & Chance, 2008, p. 162). While the definition of variability by Rossman and Chance (2008) also refers to categorical variables, no measures are presented to introductory statistics students.

In addition to these measures of variation, variability is a multi-faceted concept. Garfield and Ben-Zvi (2005) enumerated the seven key ideas related to variability:

(1) Developing intuitive ideas of variability. Variability is a characteristic of the data set. Quantities vary by differing amounts and do so for different reasons.

(2) Describing and representing variability. Different graphs of data may show different patterns of variability. In addition, standard deviation, IQR, and range can be used to describe variability in a data set, and each of these measures is more meaningful when given with a corresponding measure of center.

(3) Using variability to make comparisons. Graphs on the same scale are more easily compared. Statistics for center and spread can be used to compare graphs. There is variability both within a group and between groups.

(4) Recognizing variability in special types of distributions. The mean and standard deviation completely specify the normal distribution and can be used to estimate the proportion of data within a particular range. In a bivariate distribution, variability of both variables is important and there may be a relationship between the variables.

(5) Identifying patterns of variability in fitting models. The residuals show how well the model fits the data. Patterns might not be apparent until after data is transformed.

(6) Using variability to predict random samples or outcomes. Random samples of particular sizes have a certain amount of variability. Sample statistics vary in a particular amount based on the sample size. Variability can be predicted in chance events. The variability in samples may allow for predictions and inference.

(7) Considering variability as part of statistical thinking. Random sampling and/or comparison are necessary for inference to be used in studies. Studies should additionally try to explain variation.

In addition to these key ideas about variation, variability is not just dealt with in one manner: “statisticians sometimes attempt to minimize variability, sometimes to maximize variability, sometimes to estimate variability or simply to ‘analyze’ variance” (Gould, 2004, p. 7).

Distributions

Variation can be visually represented in the graph of a distribution. A graph of a distribution is one of several thinking tools Pfannkuch (2005) suggests for viewing variation in exploratory data analysis. Wild (2006) suggests that distributions are used as a lens through which to view variation (see Figure 1). The idea of a distribution as a lens helps to show the transformation that occurs when data is reorganized into a graph: the individual data points are often lost in the display.

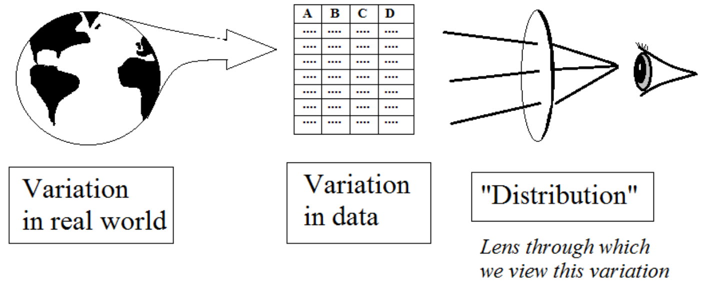


Figure 1. The concept of "distribution" regarded as a "lens" to view variation in the real world and data, adapted from Wild (2006, p. 11)

Rossman and Chance (2005) define the concept of distribution and what this means in terms of a categorical variable, "the distribution of a variable refers to its pattern of variation. With a categorical variable, distribution means the variable's possible categories and the proportion of responses in each" (p. 15). When thinking about the distribution of a quantitative variable, Konold and Pollatsek (2002) point out that the properties of a distribution including center, spread, and shape, stabilize as more data is collected when viewing "data as a mixture of signal and noise" (p. 259). This "process view better covers the range of statistical situations in which we are interested, many of which have no real population" (Konold & Pollatsek, 2002, p. 265). However, associating variability with noise in a distribution can have a negative connotation; Gould (2004) states that, "If our primary goal is to teach statistical thinking, rather than statistical techniques, then we should look to the noise, and not the signal" (p. 15).

Students' Challenges in Learning Statistics

While curriculum documents and educators have specified what statistics content and methods should be learned by students, previous research has identified some of the challenges students face when learning and reasoning about statistics. Garfield and Ahlgren (1988) found that part of the difficulty students had with learning statistics included a lack of prerequisite mathematics skills, of abstract reasoning, and of intuition for the subject. Bakker (2004) frankly stated,

> Too often, students learn statistics as a set of techniques that they do not apply sensibly. Even if they have learned to calculate mean, median, mode, and to draw histograms and box plots, they mostly do not understand that they can use a mean as a group descriptor when comparing two data sets – to give one example that is well documented. (p. 64)

Some of these difficulties stem from students' capability to do the mechanics of statistics, such as calculate the mean, median, mode, or draw a histogram, before they are ready to interpret these statistics and displays in a meaningful manner. Garfield and Ben-Zvi (2007) summarize many of the challenges of learning statistics:

> There are many misconceptions and faulty intuitions used by students and adults that are stubborn and difficult to overcome, despite even the best statistics instruction. In addition, students' statistical reasoning is often inconsistent from item to item or topic to topic, depending on the context of the problem and students' experience with the context. (p. 374)

The combination of misconceptions paired with inconsistency due to context or problem type makes statistics a difficult subject to learn.

Specific misconceptions that students have about variation and distribution have been documented through several studies. Mevarech (1983) found that students struggled with weighted averages and understanding properties of mean and variance partly due to students' view of these statistics as "just numbers" instead of characteristics of the data. The Comprehensive Assessment of Outcomes in a first Statistics Course (CAOS) test was designed to measure gains in conceptual knowledge after completion of a first college-level course in statistics (delMas et al., 2007). In a sample of 746 students, researchers found that 51.7% of students were able to rank histograms by relative size of their standard deviations at the end of the course. One common misconception was that students thought variability should be about the "bumpiness of bars" instead of the spread from the center of the distribution (delMas et al., 2007). Cooper and Shore (2010) reported that students often misinterpret histograms as if they were value bar charts (as opposed to distribution bar graphs). Batanero, Godino, Vallecillos, Green, and Holmes (1994) found that students have difficulty in choosing appropriate graphs for different situations and think of variability as unalikeability. In addition, Lem, Onghena, Verschaffel, and Van Dooren (2013) found that students often misinterpret box plots and histograms. Specifically, they may confuse histograms with bar graphs (with each bar being one observation) or might think that there is a time-scale when there is none. Design principles of Tvresky (1997) suggest that students view "higher" is "better" and that area should correspond to frequency and/or proportion, which is not the case in a boxplot (Lem et al., 2013).

Summary

This chapter reviewed the importance of statistics education for both K-12 students and college-level introductory statistics students, curriculum standards for statistics topics, the concepts of variation and distribution, and some challenges within introductory statistics content to students' learning. These topics will be of importance in this dissertation as I answer the question: how does students' reasoning about variation in a distributional context change as they progress through a college-level introductory statistics course?

Chapter 2

Literature Review

Studies focusing on students' reasoning about variation have taken place in the last 22 years since J. M. Shaughnessy's chapter in the *Handbook of Research on the Teaching and Learning of Mathematics* (1992) which reviewed all studies in mathematics education involving statistics and probability at that time. Studies in the area of reasoning about variation generally include: variation in sampling (Hjalmarson, Moore, & delMas, 2011; Noll, 2011; Noll & Shaughnessy, 2012; Noll, Shaughnessy, & Ciancetta, 2010; Peters, 2011; Peters, 2013; Reading & Shaughnessy, 2004; Sanchez, Borim da Silva, & Coutinho, 2011; Shaughnessy, 2007; Shaughnessy, Ciancetta, & Canada, 2004; Torok & Watson, 2000; Watson & Kelly, 2002; Watson & Periera-Mendoza, 1996) variation in probability (Canada, 2006; Hjalmarson, 2007; Sharma, 2007; Shaughnessy, Canada, & Ciancetta, 2003; Shaughnessy & Ciancetta, 2002; Torok, 2000; Watson & Kelly, 2004; Watson & Kelly, 2007), variation in distributions (Bakker, 2004; Ben-Zvi & Sharett-Amir, 2005; Lehrer & Schauble, 2004; Reid & Reading, 2008; Turegun, 2011; Turegun & Reeder, 2011; Watson, 2009; Watson & Kelly, 2005), and comparing variation in distributions (Ben-Zvi, 2004; Chan & Ismail, 2012; Cooper & Shore, 2008; Dabos, 2011; delMas & Lui, 2005; Inzunza, 2006; Makar & Confrey, 2005; Meletiou & Lee, 2002; Petrosino, Lehrer, & Schauble, 2003; Slauson, 2008; Watson, 2001; Watson, 2002; Watson & Shaughnessy, 2004). These studies on students' reasoning about variation have focused on elementary to graduate school students, as well as pre-service and in-service teachers. In this chapter, statistical reasoning is defined and then previous work in the different areas of variation (sampling, probability, distributions, and comparing distributions) organized by student age are discussed. After reviewing these studies, the Structure of Observed Learning Outcomes Taxonomy (SOLO), a framework used for understanding student reasoning about variation in many studies (e.g. Peters, 2011; Reading, 2004; Turegun, 2011; Watson, 2009; Watson, Callingham, & Kelly, 2007; Watson, Collis, Callingham, & Mortiz, 1995), and Hypothetical Learning Trajectories, used for predicting students' learning progress, are discussed.

Statistical Reasoning

Galotti (1989) defined informal or everyday reasoning to include evaluating arguments and choosing options. In informal reasoning, the premises are not always supplied and problems might not be self-contained. There may be multiple solutions to these problems, of varying quality, and it might not be clear whether the best solution is actually good enough. The reasoner has to figure out what information is relevant.

Statistical reasoning can be found in "stages in people's thinking where they are asked to state implications, justify a conclusion, or make an inference" (delMas, 2005, p. 85) and that "statistical reasoning is demonstrated when a person can explain why a particular result is expected or has occurred, or explain why it is appropriate to select a particular model or representation" (p. 85). People are typically able to reason more effectively about a topic when they are familiar with it, although this can also lead to the inclusion of irrelevant personal knowledge in their reasoning.

Variation in Sampling

Much of the research on students' reasoning about variation in sampling situations stems from the Lollie Task (Shaughnessy, 2007). Sampling tasks include those where students either predict or take repeated samples from some population. This section will address findings on the Lollie Task, the main item used to assess student reasoning in this area, and other tasks by tracing research through various student age groups from elementary school students to teachers.

Elementary school students. Two studies mainly focused on elementary school students' reasoning. Watson & Periera-Mendoza (1996) found that when 3rd through 6th grade students in Australia and Canada were asked questions about bar graphs and making predictions, students' real-world knowledge often inhibited their ability to make statistical predictions. In order to further students' understanding, researchers recommended asking questions that force students to consider changes in the graph due to chance. Watson and Kelly (2002) found that third grade students (N = 72) improved from a pretest to a posttest ($p < 0.001$) mainly focusing on variation in sampling after completing a ten-lesson unit. The unit included many different situations such as packages of colored candies, families, and times students could stand on one foot and topics including sampling, chance, and dot plots.

Middle school and high school students. Many of the studies focused on middle school and high school students' reasoning about variation in sampling stem from the Lollie Task (Reading & Shaughnessy, 2004; Noll & Shaughnessy, 2012; Shaughnessy, 2007; Shaughnessy, Ciancetta, & Canada, 2004; Torok & Watson, 2000). The task was developed from the "gum ball" question on the 1996 NAEP exam that focused on centers of the distribution of samples. Reading and Shaughnessy (2004) reformulated the task in order to highlight students' reasoning about variation in the sampling situation (see Figure 2).

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

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

1c) Look at the possibilities that some students have written down for the numbers they through likely. Which one of these lists do you think best describes what might happen? Circle it.
a) 5,9,7,6,8,7
b) 3,7,5,8,5,4
c) 5,5,5,5,5,5
d) 2,3,4,3,4,4
e) 7,7,7,7,7,7
f) 3,0,9,2,8,5
g) 10,10,10,10,10,10
Why do you think the list you chose best describes what might happen?

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

Figure 2. Lollie Task student response form adapted from Reading & Shaughnessy (2004, p. 212). This exercise was used to assess students' reasoning about variation in a sampling situation.

Shaughnessy, Ciancetta, & Canada (2004) found that middle school and high school students used additive (more of a particular color), proportional (based on the population proportion), and distributional reasoning (taking into account both center and spread) on the Lollie Task. In addition, researchers found that students describe variability in ways that are too high (most predictions above the population proportion), too low (most predictions are below the population proportion), too wide (predictions cover the entire range of possibilities, not just likely responses), or too narrow (little to no variability in predictions). After physically doing the task, more students provided reasonable answers based on computer simulations of the distribution (Shaughnessy, 2007).

The Lollie Task evolved to ask students to additionally draw a histogram of what the sampling distribution might look like (Noll & Shaughnessy, 2012). Noll and Shaughnessy (2012) reported teaching episodes on variability as well as the survey and task-based interviews in middle schools and high schools in both urban and rural locations. It was noted that the middle school group from a mathematics and science magnet school performed nearly as well as the AP Statistics student group. In addition, researchers found that students' prediction ranges did not necessarily match up with their graphical displays (Noll & Shaughnessy, 2012).

The Lollie Task has also provided data to inform the Hierarchy of Reasoning about Variation, which consists of two hierarchies, the Description Hierarchy and the Causation Hierarchy (See Table 2).

Table 2

Hierarchy of Reasoning about Variation adapted from Reading & Shaughnessy (2004, pp. 214–219). This hierarchy was used to assess students' reasoning in the Lollie Task.

Description Hierarchy	Causation Hierarchy
D1- Concern with either middle values or extreme values	C1- Identify extraneous causes of variation
D2- Concern with both middle values and extreme values	C2- Discuss frequencies of color(s) as cause of variation
D3- Discuss deviations from an anchor (not necessarily central)	C3- Discuss proportion(s) of colors as the cause of variation
D4- Discuss deviations from a central anchor	C4- Discuss likelihoods based on proportions

Although the tasks did not ask students to explain possible causes of the variation, students felt obligated to mention the cause (Reading & Shaughnessy, 2004). The Causation Hierarchy focused only on variation due to sampling; however, many other sources of variation exist. Wild (2006) categorized the sources of variation in data in the following way (see Figure 3):

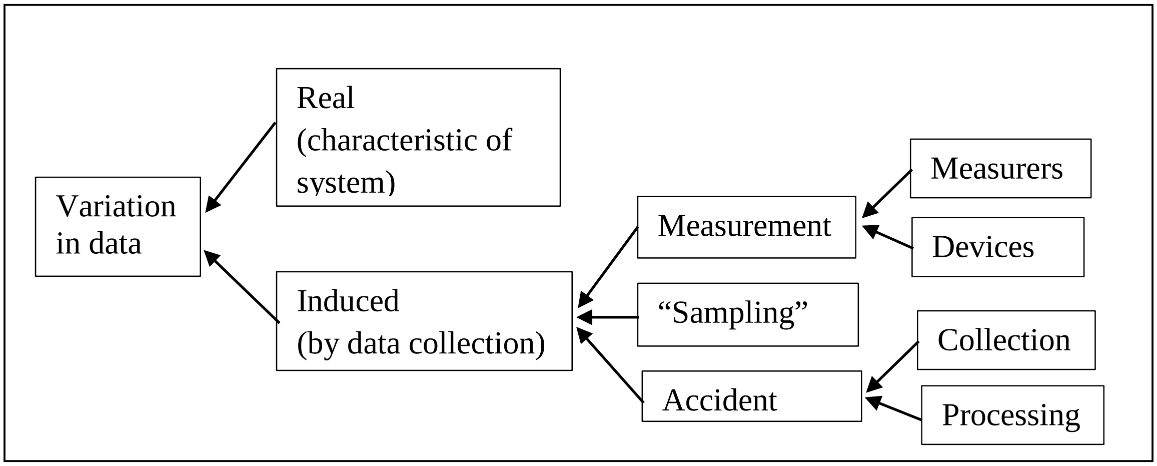


Figure 3. Diagram explaining the sources and types of variation in data adapted from Wild (2006, p. 12)

Depending on the nature of the task, some of these sources may be present while others are not.

A second framework (Figure 4) on students' reasoning about variation in a sampling situation also originated from student work on the Lollie Task with middle school and high school students (Noll & Shaughnessy, 2012). Answers coded "other" did not supply reasoning or were unclear. Answers coded "additive" were based on there being more than one category, but not proportional reasoning. Answers coded "weak center" contained evidence of using the mode. Answers coded "shape" made mention of the distribution's shape. Answers coded "variation" contained a measure of variation or referred to the tails of the distribution. Answers coded "median, mean, proportional" based the center of the distribution on the population proportion. Lastly, answers coded "distributional" coordinated at least two of the following: shape, center, and variation.

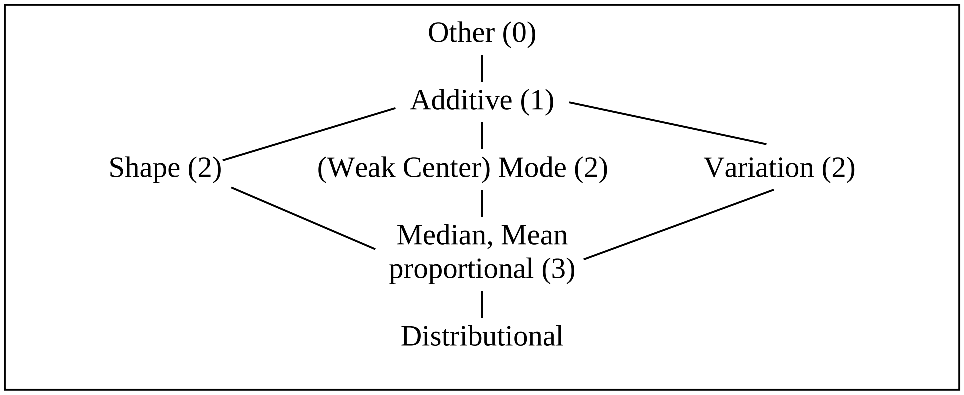


Figure 4. Conceptual lattice of students' reasoning about variation from the Lollie Task adapted from Noll & Shaughnessy (2012, p. 523)

Torok and Watson (2000) interviewed students (*N*=16) to better understand their reasoning on the Lollie Task and two real-world situations: daily maximum temperature in their city and heights of upper-primary-level students. The study produced a framework for categorizing the level of reasoning about variation displayed by students in Grades 4, 6, 8, and 10. The framework consists of four levels of student reasoning: weak appreciation of variation, isolated appreciation of aspects of variation and clustering, inconsistent appreciation of variation and clustering, and good, consistent appreciation of variation and clustering. Students in 4th and 6th grade responded in the first two levels, while the 8th grade students responded at the second (3 students) and fourth levels (1 student), and the 10th grade students responded at the highest two levels (Torok & Watson, 2000).

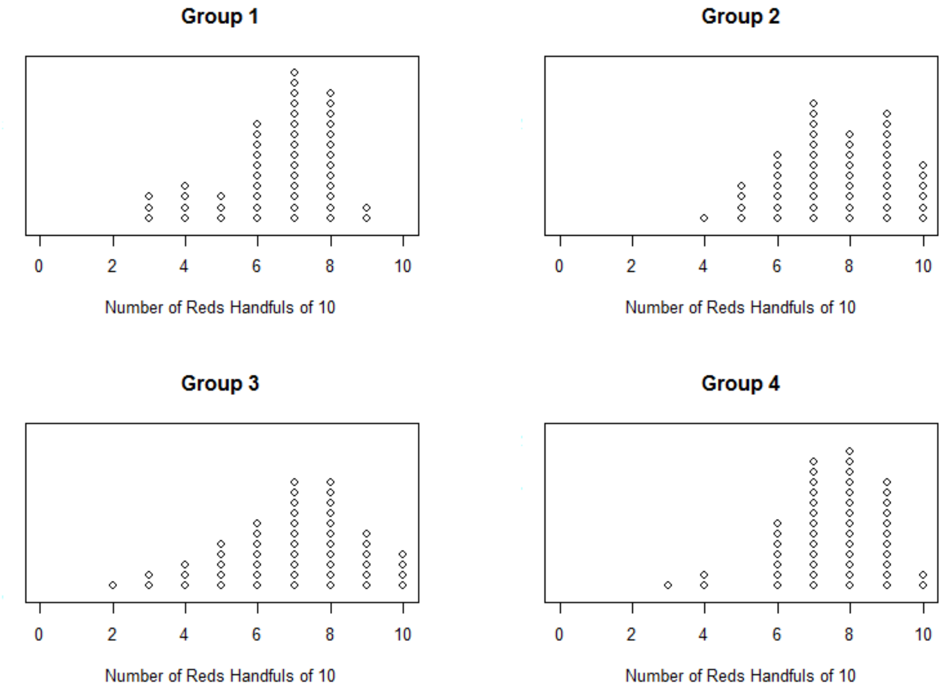
College students. Hjalmarson, Moore, and delMas (2011) looked at how students quantify variation in an undergraduate engineering course where students had "to develop a procedure for quantifying the roughness of a surface at nanoscale" (p. 15). The data were given in terms of a picture of the surface with different colors indicating height of the surface. Researchers found that only four out of the 35 procedures were able to be used on another sample or needed very minor revision. In addition, researchers analyzed the statistical measures students used and found that 23 out of 35 procedures utilized standard deviation while only nine used other measures of spread.

Teachers and instructors. Sanchez, Borim da Silva, and Coutinho (2011) reviewed literature on teachers' understanding of variation and found that most studies treated teachers and students' understanding in the same way, although teachers likely need an understanding of statistics for teaching, akin to mathematical knowledge for teaching.

The Lollie Task has also been used with college-level statistics teaching assistants in a modified format. It was found that although the teaching assistants (some of whom had taken many graduate courses in statistics) might be able to identify the formal distribution associated with the problem, many had the same misconceptions as the students they were teaching such as predicting too large of a range (Noll, 2011).

A second task stemming from the Lollie Task, the Real/Fake Task, has been used with students from middle school to graduate teaching assistants (Noll, Shaughnessy, & Ciancetta, 2010). The Real/Fake Task (see Figure 5) requires students to decide which empirical sampling distributions are real and which are fake (note: Graphs 1 and 3 are fake, Graphs 2 and 4 are real). The middle school research group had the largest percentage (35%) of students getting all four identifications correct, while 30% of the high school research students and 28% of graduate students made four correct identifications. However, in the reasoning of students, which were categorized as not applicable, shape, center, spread, tails, or multiple reasons, around 50% of undergraduate students and graduate students used shape while only 16% to 42% of the middle school and high school student groups used shape. Lastly, the percentage of students who referenced two or more of the categories tended to increase with age.

A class conducted an experiment, pulling 50 samples of 10 candies from a jar with 750 reds and 250 yellows, and graphed the number of reds. However, in this class some of the groups "cheated" and did not really do the experiment, they just made up a graph. Here are some of the students' graphs from that class.



A) Which graphs do you think are real? ___________ Explain the reasons for your choices.
B) Which graphs do you think are made-up? __________ Explain the reasons for your choices.

Figure 5. Real/fake task handout based on the Lollie Task situation adapted from Noll, Shaughnessy, & Ciancetta (2010, p. 2)

Peters (2011) studied AP statistics teachers' robust understanding of variation which she described as "integrated reasoning about variation within each perspective and across perspectives for four elements: variational disposition, variability in data for contextual variables, variability in relationships among data and variables, and effects of sample size on variability" (p. 52) using semi-structured content interviews. Furthermore, Peters (2013) also studied five AP statistics teachers who were identified as having a robust understanding of variability, it was recommended that a course focusing on Exploratory Data Analysis (EDA) beginning with the design of experiments would enhance understanding for teachers. In addition, these teachers found that active learning and simulations were useful in extending their understanding of variability in design, data, and modeling.

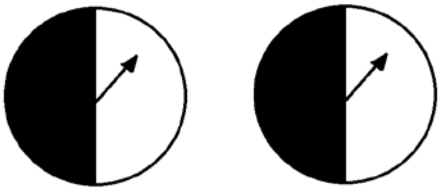
Variation in Probability Context

Studies exploring students' reasoning about variation in the probability context are rather diverse in terms of types of student and types of questions and often involved other contexts as well. Two tasks that have frequently been used in this area include the Spinner Task and the Die Task. In the spinner task, students are asked to guess a typical sequence of outcomes of a half black, half white spinner when completing six sets of 50 spins. In the Die Task students are asked to guess a typical distribution of 60 rolls of a six-sided die. Watson and Kelly (2007) provided rubrics for teachers to assess these tasks, as well as the Lollie Task. In the probability context, they note that formal probability training influences students to focus on expectation (rather than variation) when making predictions, although students first acknowledge variation before expectation. In addition, Torok (2000) gave a practical implementation of the Spinner Task and Lollie Task in classroom and found that students were able to explain variation in their own (i.e., non-standard) language.

Elementary and middle school students. Watson and Kelly (2004) studied student responses regarding variation in a probability context as well as reasoning about a distribution. The researchers wanted to test (1) whether student scores improved after instruction and whether this improvement was sustained two years later, (2) whether students acknowledge the role of variation when predicting the outcomes for a single and/or repeated trial of a spinner, and (3) whether students could accurately identify appropriate and inappropriate variation in a given graphical display of a distribution. The study was longitudinal and included a pre-survey, teaching intervention, post-survey, and then a follow-up survey two years later. The surveys in all three cases were identical and questions focused on a spinner that was half black and half white and the number of spins out of 10 or 50 that one might get black. On the variation point estimate scale, fifth grade students (n = 58) improved from a mean of 4.55 to a mean of 5.94 ($p < 0.002$), seventh grade students (n = 66) had no statistically significant change, and ninth grade students (n = 28) improved from a mean of 5.79 to 6.76 ($p < 0.03$) two years later on a ten-point scale (Watson & Kelly, 2004).

Shaughnessy, Canada, & Ciancetta (2003) found middle school students were less likely to acknowledge variability in a probability environment versus a sampling environment. They used the Lollie, Spinner, and Die Tasks. Over half of students predicted no variation on the die task, which researchers partially attributed to prior classroom instruction typically focusing on calculating probabilities of single events with die.

High school students. Shaughnessy and Ciancetta (2002) gave a survey containing a different spinner question from the NAEP using two half black and half white spinners to students taking mathematics courses in grades 6 through 12 (N = 652). See Figure 6. In the follow-up interviews with 28 participants, which included students physically using two spinners to try the task, researchers found that students' scores improved and more students had correct reasoning after experimenting during the interview. In addition, the successful students tended to list the sample space.



The two fair spinners above are part of a carnival game. A player wins a prize only when *both* arrows land on black after each spinner has been spun once.

Jeff thinks he has a 50-50 chance of winning. Do you agree?

A. Yes B. No Justify your answer.

Figure 6. NAEP spinner task adapted from Shaughnessy & Ciancetta (2002, p. 2) used with middle school and high school students before and after experiment with spinners.

College students and pre-service teachers. First-year engineering student teams were given an open-ended problem to rank shipping companies based on their likelihood of an on time arrival. Student responses were classified by the measures of center and variability used. One feature of the data set was that all of the means were similar, so students would likely need to use other measures as well in their ranking procedure. Out of 51 teams, 92% used mean, 27% used mode, and 25% used median (teams could use more than one measure) as part of their ranking procedure. Large numbers of groups also used measures of variability, though fewer than used measures of center: 78% used standard deviation, 49% used the maximum value, and 25% used the minimum value as part of their ranking procedure (Hjalmarson, 2007).

Canada (2006) studied how elementary pre-service teachers' understanding of variation in a probability context changed after a teaching intervention in a mixed-methods study. Canada found an evolving framework for the teachers' understanding of variation. This framework included three aspects of variation: expecting, displaying, and interpreting variation. Expecting variation included the dimensions that describe what is expected and why it is expected. Displaying variation included the dimensions that produce graphs, and evaluate and compare graphs. Interpreting variation included the dimension of the causes and effects of variation and the dimension that influences expectations and variation. In addition to this framework, students gained an appreciation for variation after the in-class interventions (hands-on activities, computer simulations, and discussions) and showed a statistically significantly greater use of proportional reasoning on two questions (Canada, 2006).

Sharma (2007) worked with pre-service teachers in Australia and studied their responses to two questions: one on the spread of sampling distributions and one on the Die Task. Reponses were analyzed on a scale of being statistical, partial-statistical, or non-statistical. It was found that seven out of 24 students were able to give statistical answers to the sampling distribution problem while only two out of 24 students gave statistical answers to the Die Task. Only two students gave non-statistical answers to the sampling distribution problem while four students gave non-statistical answers to the die task.

Variation when Reasoning about a Distribution

Studies involving reasoning about variation in a distribution have considerable crossover to and have utilized a much wider range of tasks than other areas of reasoning about variation. Researchers have focused on the understanding of elementary and middle school students, as well as college-level students.

Elementary and middle school students. In a case study with three second grade students working on an open-ended task, researchers found that students tended to make lists or a table instead of bar graphs. One of the students was able to display the idea of probable values and frequencies, while the other two students were only able to come up with probable values (Ben-Zvi & Sharett-Amir, 2005).

In a hands-on activity, a 5th grade class collected real data through plant growth experiments they designed. One part of this project focused on displaying the heights of the plants, first with their own creative methods, and later with histograms. Students were interviewed to assess their understanding. It was found that students preferred displays that did not lose information (i.e., stem and leaf plots versus histograms). Furthermore, students struggled with understanding quartiles, probability, and that there was measurement error as well as natural variation among the plants (Lehrer & Schauble, 2004).

A large scale study by Watson, Kelly, Callingham, and Shaughnessy (2003) extensively surveyed 746 students in grades 3, 5, 7, and 9. Topics on the survey included probability, sampling, distributions, and measures of spread. Data were analyzed using a Rasch model technique. The questionnaire was designed to be used to measure outcomes from teaching interventions. Researchers found four levels of understanding of variation:

pre-requisite knowledge, partial understanding, application of understanding, and complex reasoning (2003).

Watson and Kelly (2005) found a plateau in performance in grade 9 students' ability to explain variation in weather data, while students in grades 3, 5, and 7 showed an increase in ability. Moreover, the ability to graph a situation and describe what variation and variable mean increased with grade level. Watson and Kelly recommended an increase of writing in mathematics, which would enable students to better describe the relationship between center and variability in a distribution.

Watson (2009) studied students' integration of reasoning about variation in the context of reasoning about a distribution. Watson interviewed 109 students from ages 6 to 15 in increments of 2 to 3 grades. During the interviews, students were asked to make graphical representations of three different data sets of varying degrees of difficulty. With the books data set, participants were asked to represent the number of books certain people had read from a verbal description. In the weather data set, students were asked to graph the average maximum temperature over a year if the yearly average was 17 degrees Celsius. Finally, in the Lollie Task, as described previously, participants were asked to make a graph of the number of red lollies in 40 samples of size 10. During the interviews, participants were also asked questions about their representations and to make predictions. Watson (2009) found that students acknowledged variation before they acknowledged expectation in a distribution.

Bakker (2004) conducted a classroom-based study with 8th grade Dutch students to investigate their progress as they learned to reason about the shape of a distribution. Students "grew" a sample of weights of 8th graders: first a sample of size 10, then a sample of size 27, followed by a sample of size 67, and finally a sample of the whole city. Their samples were also compared to real data from another school. Students were able to not just focus on individual data points, but also to reason about a continuous shape of the distribution for an entire city. Bakker suggested that whole-class discussions and individual questioning by interviewers in the classroom helped students to make the most progress in their ability to reason about the shape of a distribution.

College students. Reid and Reading (2008) further refined their consideration of variation hierarchy (initially developed from the Lollie Task by Reading and Shaughnessy in 2004) using test questions and a homework assignment in a college-level statistics course. This hierarchy (Figure 7) could be used to code tasks to understand students' reasoning about variation in sampling, probability, distribution, and comparing distribution situations. It contains four levels: no consideration of variation, weak consideration of variation, developing consideration of variation, and strong consideration of variation, along with the types of answers that fit into these levels.

Turegun (2011) focused on students' conceptual understanding of range, interquartile range, and standard deviation, especially their pre-existing conceptions, how they articulate their understanding, and connections between measures of spread and distributions. Turegun (2011) found that although students' (n=29) reported that they guessed on the CAOS Measures of Spread pretest, their scores were statistically significantly higher than guessing on this multiple choice test (p < .01). On four questions related to standard deviation, over half of the students answered correctly on the pretest. In addition, Turegun and Reeder (2011) found a lack of student understanding of histograms.

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

Figure 7. A generalized Consideration of Variation Hierarchy adapted from Reid & Reading (2008) used to assess student responses to all areas of reasoning about variability.

Variation when Reasoning about Comparing Distributions

When students are asked to compare variation in distributions, they have to understand the concept of distribution and also be able to distinguish between different distributions. Students as young as fourth grade (Petrosino et al., 2003) to college-level instructors (Dabos, 2011) have served as participants in studies focusing on reasoning when comparing distributions.

Elementary and middle school students. In a teaching experiment with fourth grade students, Petrosino, Lehrer, and Schauble (2003) found that ten out of 14 students in one-on-one interviews were able to see the difference in variability between bimodal and unimodal distributions of similar ranges by using *spread numbers*, which is the average distance of data from median. Students measured the heights of rockets with round or pointed tips, and then used their collected data to reason about which rocket might go higher (a confliction with their hypothesis).

In a case study of two Israeli Grade 7 students, Ben-Zvi (2004) found the following progression in an activity focused on comparing the length of American and Israeli student last names where both data sets were of equal size (see Table 3).

Table 3

Student progression of two grade 7 students on reasoning about variability on a task comparing the lengths of last names of American and Israeli students (Ben-Zvi, 2004, p. 48)

Stage 1. On what to focus: Beginning from irrelevant and local information
Stage 2. How to describe variability informally in raw data
Stage 3. How to formulate a statistical hypothesis that accounts for variability
Stage 4. How to account for variability when comparing groups using frequency tables
Stage 5. How to use center and spread measures to compare groups
Stage 6. How to model variability informally through handling outlying values
Stage 7. How to notice and distinguish the variability within and between the distributions in a graph

This progression focused on students' reasoning beginning with individual data points to being able to compare variability within and between the distributions.

A longitudinal study by Watson (2001) focused on students' ability to graphically compare two data sets (Figure 8) and describe which group did better on a test. The first three comparisons all used the same frequency in both groups, and in the last, students had to deal with two different sized groups. Students in Grades 3, 5, 7, and 9 participated in a follow up study three or four years later, in which the same interview protocol was used. It was found that 62% of students improved, 5% stayed at the highest level, 24% stayed at the same level (non-optimal), and 10% performed at a lower level. Due to the sampling plan, researchers attributed the improvement to students' general development, and not to the curriculum. A second study (Watson, 2002) that used the same task introduced cognitive conflict by showing participants videotapes of other students' reasoning. After watching the videos, 57% of students improved their answers on part c (of Figure 8) while 30% of students improved their answers on part d; however, fewer students discussed the variation present in the graphs. In a further study utilizing this same task, Watson & Shaughnessy (2004) studied students in grades 3 to 13. Younger students tended to use total numbers of students with particular scores to make decisions about which class did better on the test. This method was not appropriate for the last pair of distributions (part d) due to the difference in frequency in the two classes. It was also noted that some students used the mean to make a decision on this task and a few students used the shape of the distributions (one was normal, one was skewed left) in order to justify their answers. Researchers recommended the need to explicitly connect proportional reasoning with data and chance.

High school students. In Chan and Ismail (2012), 412 tenth grade Malaysian students from nine secondary schools responded to a question asking students to decide which of two histograms had a larger standard deviation. One distribution was uniform, with a larger range, while the other distribution was mound-shaped. Two misconceptions about variability were identified: that the mean and the standard deviation are the same and that the frequency is the same as the standard deviation. Researchers recommended

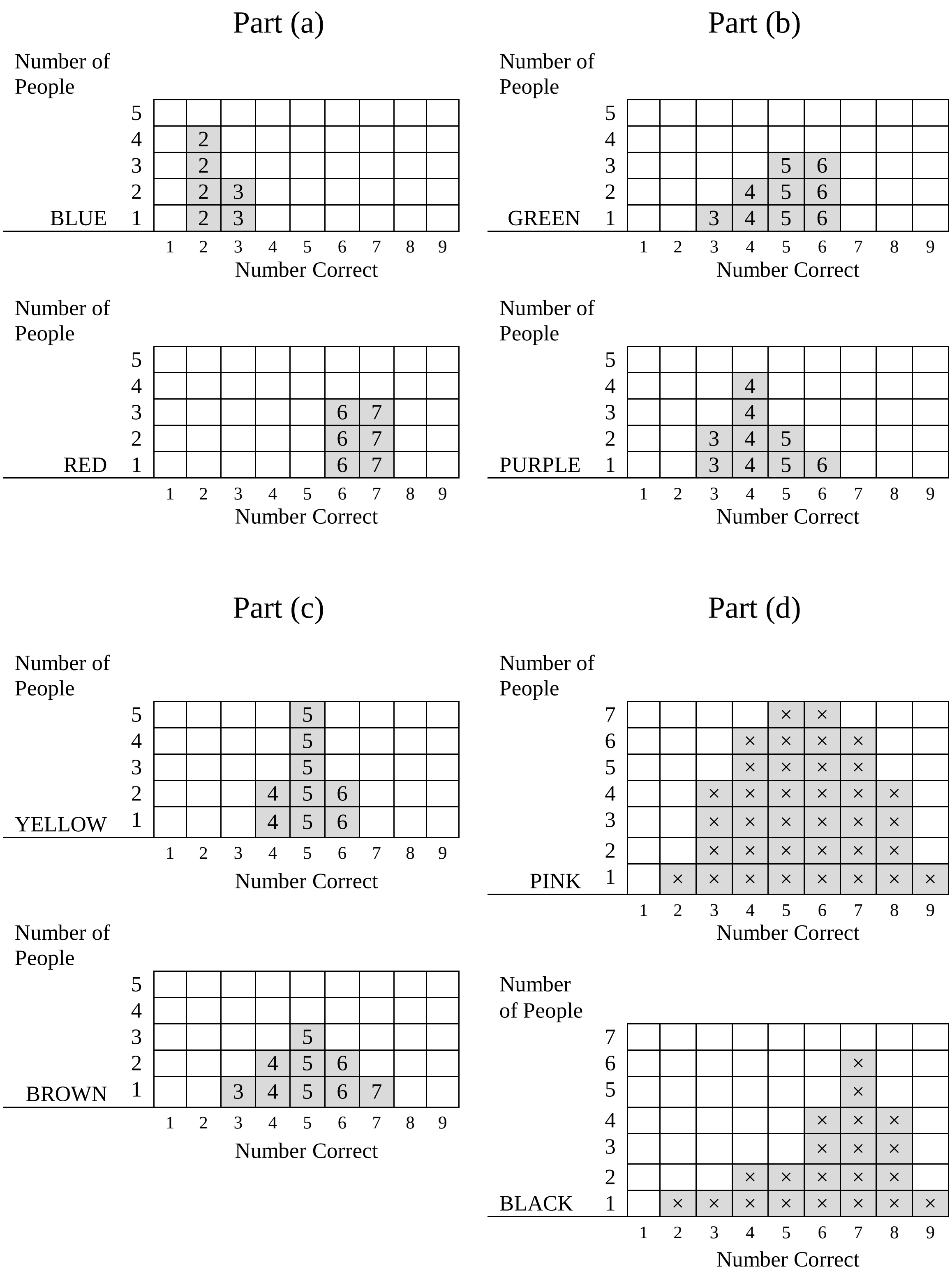


Figure 8. Four pairs of graphs adapted from Watson (2001), Watson (2002), and Watson and Shaughnessy (2004). Students were asked to identify the class that did better on the test for each pair of graphs.

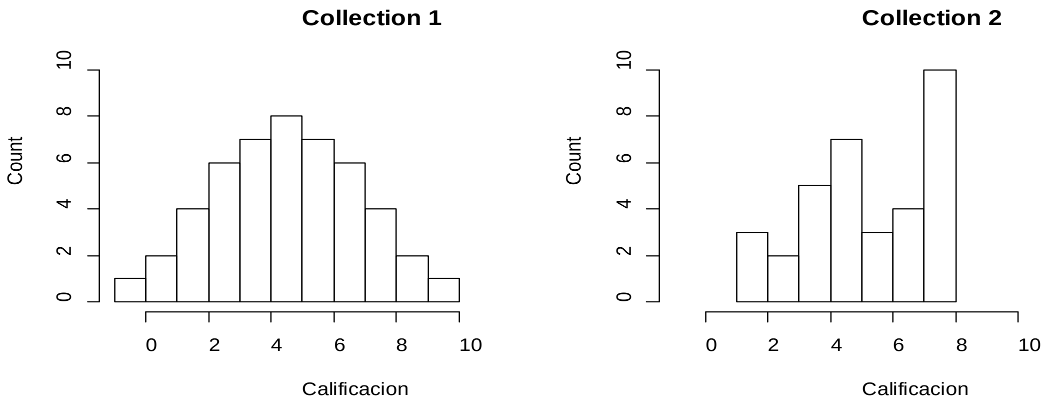
that students use technology to better understand the concept of standard deviation.

College students and pre-service teachers. At the introductory college level, delMas and Liu (2005) conducted an exploratory study of students' conceptions of standard deviation using an interactive computer environment and histograms. The twelve participants were interviewed while working through five "games" of moving the bars of histograms around to find the arrangements with the largest and smallest standard deviations possible in order to understand why they chose the arrangements. After a period of exploration with the games, participants took a ten-question test that asked them to choose between two histograms. The interviewer also asked the participants to explain why they chose each answer. The computer program gave students feedback as to whether they were correct after each question on both the exploration phase and the test. On the test, nine students got 9 out of 10 questions correct and three students got 7 out of 10 questions correct. Researchers found the following ways that students would use to explain which histogram had a larger/smaller standard deviation that indicated a good understanding of standard deviation: far away-mean, balance, more values in the middle, and bell-shaped rules. Other students showed a developing sense of reasoning about standard deviation with the following explanations: contiguous, range, mean in the middle, and far away-values (delMas & Liu, 2005).

Cooper and Shore (2008) conducted a study that identified student misconceptions about center and variability in histograms and stem-and-leaf plots across three different levels of a first college statistics course: a non-calculus based statistics class for elementary education majors, a different non-calculus based statistics class for non-education majors, and an introductory calculus-based statistics class. The misconceptions included that bell-shaped histograms with more variable bars had greater variability (approximately 27% of students) and that any two histograms with the same range had the same amount of variability regardless of other features of the distribution (approximately 20% of students). Students in higher level statistics courses did not perform significantly better on a four-item assessment designed by the researchers than the students in the lower level courses. The researchers recommended the use of histograms to help students learn to make “valid comparisons between shape and relative variability” (p. 11). These results echo those of Meletiou and Lee (2002), who found that understanding histograms is important to understanding variability even when the college-level course focused on concepts and doing simulations. Meletiou and Lee (2002) also found that students were often confused as to whether they should look at the x- or y-axis in order to reason about variability.

Inzunza (2006) conducted interviews with 11 students aged 19 to 21 who were asked to compare two histograms (see Figure 9). In these histograms, there are 48 data points in collection 1, and 34 data points in collection 2; however, both collections have an IQR of 4. Student misconceptions included: “variability depends on the quantity of data” and “variability depends on also the irregularity of the distribution” (p. 246). Furthermore, Izunza found that students had a lack of understanding of standard deviation and how it relates to empirical distributions.

Slauson (2008) used the predict/test/evaluate lab model to create two activities focusing on standard deviation and standard error. The model worked well for increasing students’ understanding of standard deviation and margin of error in a confidence interval, but not as well for increasing students’ understanding of standard error and how



1. Mark with an X the distribution which has more variability and explain with details the reasons of your election.

Figure 9. Histograms and question adapted from Inzunza (2006, p. 245)

it differs from standard deviation. Slauson (2008) found no statistically significant differences on CAOS test scores, however, not all of the test questions focused on these topics, so only small differences were expected.

In order to better understand the language of variation, pre-service teachers were asked to compare two dot plots with different numbers of data points in a study conducted by Makar and Confrey (2005). Students referred to spread by using terms such as: spread out, scattered, evenly distributed, or dispersed. Variation-talk by students typically fell into four types: spread, low-middle-high, modal clump, and distribution chunks. The percentage of students that mentioned spread increased by 24% (*n* = 17) from a test at the beginning of the semester to the end of the semester.

Teachers and instructors. Dabos (2011) studied community college instructors' conceptions of variation. Some of the participants in the study had taught statistics and some had degrees in statistics in addition to mathematics. The survey and interview protocols included the Lollie Task, Die Task, Real/Fake Task, and other questions comparing dot plots and histograms. It was found that instructors' knowledge of probability affected their reasoning on the Lollie Task. In addition, instructors had difficulty comparing two distributions if unable to use range: only 48% answered these questions correctly.

Structure of Observed Learning Outcomes (SOLO) Taxonomy

One general model frequently applied to understanding reasoning about variation is the SOLO Taxonomy, which was developed by Biggs and Collis in 1982 for non-subject specific use (see Table 4). In this model, Piaget's developmental stages are contrasted with the SOLO levels. In Table 4, *Capacity* refers to the attention span that the level of SOLO requires, *relating operation* is how the cue and response connect, and *consistency and closure* refers to the internal struggle between coming to a conclusion and giving a consistent response. One of the main differences between Piaget's learning theory of stages and the SOLO model is that the SOLO model is only used to classify the response, not the student.

In 1991, the SOLO taxonomy was refined by Biggs and Collis to a neo-Piagetian model with five main modes of reasoning: Sensori motor, Ikonic, Concrete symbolic, Formal, and Post-formal. Within each mode are cycles: Uni-structural (considering one relevant aspect), Multi-structural (considering several disjoint but relevant aspects), and Relational (integration of several aspects). Pre-structural responses are contained within the previous mode, since the students are not able to show meaningful understanding, and the extended abstract responses are really in the next higher mode, due to the nature of the generalization they are able to make. See Figure 10.

Table 4

SOLO Model adapted from Biggs and Collis (1982, pg. 24-25). This model has been used to assess the quality of reasoning of participants in many areas including reasoning about variability

Developmental base stage with minimal age	SOLO description	Capacity	Relating operation	Consistency and closure
Formal Operations (16 + years)	Extended abstract	Maximal: cue + relevant data + interrelations + hypotheses	Deduction and induction. Can generalize to situations not experienced	Inconsistencies resolved. No felt need to give closed decisions – conclusions held open, or qualified to allow logically possible alternatives. (R1, R2, or R3)
Concrete generalization (13-15 years)	Relational	High: cue + relevant data + interrelations	Induction. Can generalize within given or experienced context using related aspects	No inconsistency within the given system, but since closure is unique so inconsistencies may occur when he goes outside the system
Middle Concrete (10-12 years)	Multi-structural	Medium: cue + isolated relevant data	Can "generalize" in terms of a few limited and independent aspects	Although has a feeling for consistency can be inconsistent because closes too soon on basis of isolated fixations on data, and so can come to different conclusions with same data
Early Concrete (7-9 years)	Uni-structural	Low: cue + one relevant datum	Can "generalize" only in terms of one aspect	No felt need for consistency, thus closes too quickly: jumps to conclusions on one aspect, and so can be very inconsistent.
Pre-operational (4-6 years)	Pre-structural	Minimal: cue and response confused	Denial, tautology, transduction. Bound to specifics	No felt need for consistency. Closes without even seeing the problem.

The SOLO Model has been used by researchers in statistics education to analyze student responses in situations involving reasoning about variation. Some researchers focused only on using the modes from the 1982 model (Turegun, 2011; Watson, 2009;Watson, Callingham, & Kelly, 2007) or several cycles within a subset of modes

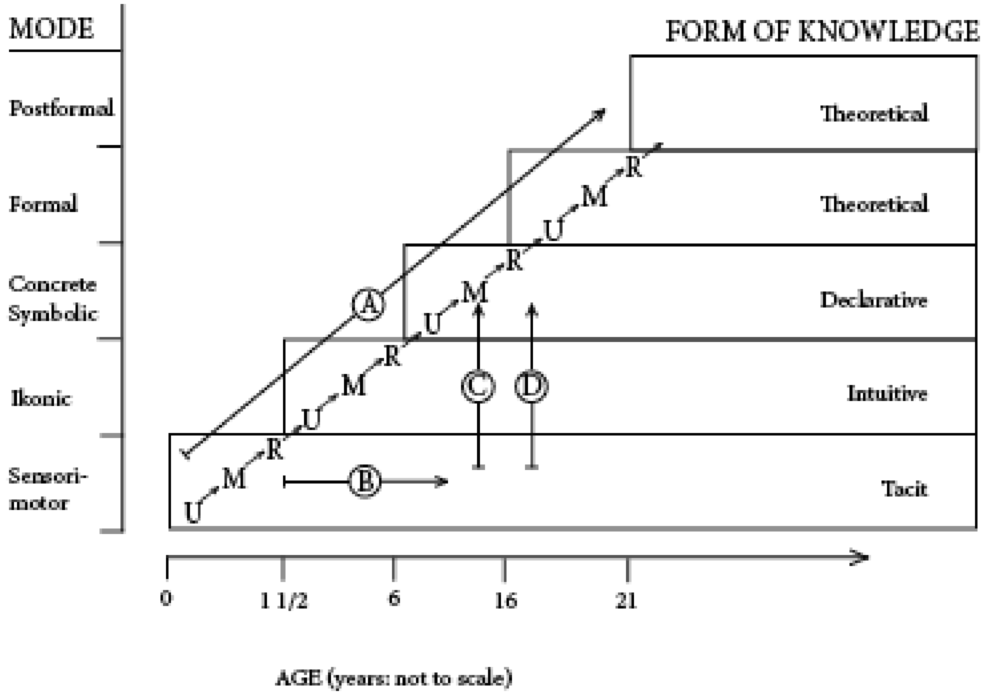


Figure 10. The relationship of modes, cycles, and forms of knowledge in the SOLO model, adapted from Biggs and Collis (1991, p. 66).

from the 1991 model (Peters, 2011; Reading, 2004; Watson, Collis, Callingham, & Mortiz, 1995) to classify student responses.

SOLO modes. Watson (2009) used the SOLO model to classify responses to tasks where students made or described distributions using the books data, weather data, and Lollie Task, in the follow way:

Table 5

SOLO model responses for descriptions of variation in distributions, adapted from Watson (2009, p. 38).

Level 0: Idiosyncratic- No indication of variation or expectation Level 1: Unstructured variation Level 2: Variation shown by value Level 3: Initial acknowledgement of expectation Level 4: Integration of variation and expectation

For example in the Lollie Task, when students were asked to draw the number of red lollies in 50 handfuls, a response at level 0 was a picture of the bowl of candies. A response at level 1 was a short list of possible outcomes. A response at level 2 was a time plot with too much variation, while a response at level 3 was a time plot with appropriate variation. A response at level 4 was a histogram with appropriate variability and center (Watson, 2009).

Watson, Callingham, and Kelly (2007) interviewed students from kindergarten to Grade 9 on six tasks (Lollie, weather, comparing groups, Spinner, comparing population and sample, definition of variation) used the five levels of the SOLO model as well as the Rasch Partial Credit Model to define six levels of understanding and place students in them. These levels were defined in Table 6.

Table 6

SOLO modes and responses for describing variability in six different distributions, adapted from Watson, Callingham, and Kelly (2007, p. 93)

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

SOLO cycles. Watson, Collis, Callingham, and Mortiz (1995) used the SOLO model to assess higher-order thinking in a situation where six Grade 6 and one Grade 9 students were given 16 cards with information about students on them and were asked what kinds of questions they could ask with the cards. Multiple uni-structural, multi-structural, relational cycles were identified as students learned about a concept.

Reading (2004) made a refinement of the description of variation hierarchy (Reading & Shaughnessy, 2004) using the SOLO model that took into account qualitative (described center/variation) and quantitative (used measures of variation) responses where the qualitative responses were considered less sophisticated. Students in Grades 7, 9, and 11, were asked to evaluate whether a particular month would be appropriate to hold an outdoor event based on weather and temperature data. Two cycles of uni-structural, multi-structural, and relational within the concrete symbolic mode were identified in student responses. The first cycle consisted of qualitative responses, while the second cycle consisted of the higher-order quantitative responses.

Lastly, the SOLO model has also been used to describe even more sophisticated levels of reasoning. Peters (2011) used the SOLO model to describe the highest levels of reasoning about variation in her work with AP Statistics teachers. Peters' framework is given in Figure 11. This framework classifies teachers' reasoning by design, data-centric, and modeling viewpoints and shows that the integration of these viewpoints is the second cycle of SOLO levels.

Hypothetical Learning Trajectory

A hypothetical learning trajectory is "the teacher's prediction as to the path by which learning might proceed" (Simon, 1995, p. 135). When learning particular concepts, there are often multiple ideas that need to be understood to be able to get to the goal of the lesson so the hypothetical learning trajectory predicts the way to get from one idea to another to the overall goal. There are three components in a hypothetical learning

Elements and Reasoning Indicative of Robust Understanding of Variation

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

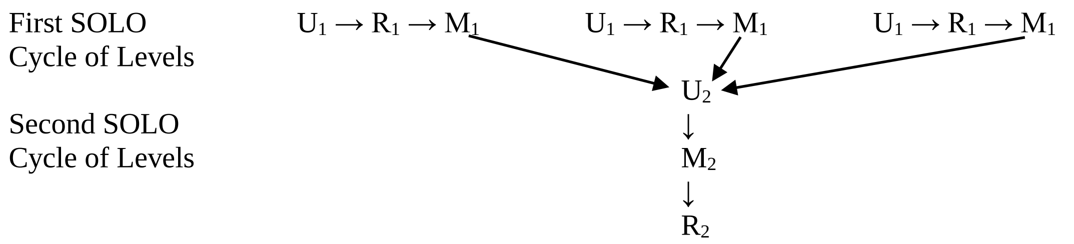


Figure 11. SOLO model describing AP statistics teachers' variational reasoning from design, data-centric, and modeling perspectives, adapted from Peters (2011)

trajectory: the direction that is defined by the learning goal, the learning activities, and the prediction of how students will reach the learning goal. This model helps those with a constructivist perspective to better understand how mathematics should be taught (Simon, 1995). Two studies relevant to students' reasoning about variation have utilized a hypothetical learning trajectory.

Garfield, delMas, and Chance (2007) conducted two Japanese lesson studies on students' reasoning about variability. They found that the largest student gains were made at the end of the course, long after all these topics had been taught. However, at the end of the course, few students used the expected measures of variability (standard deviation and IQR), while many used the range and the shape of the distribution in order to reason about variability. The authors noted that it was difficult to find a problem that was truly relevant for this topic, thus they only used data that was relevant to students. Their hypothetical learning trajectory is described in Table 7.

Table 7

Hypothetical Learning Trajectory for reasoning about variability, adapted from Garfield, delMas, and Chance (2007, p. 142)

- Begin with students' basic understanding that data vary
- Investigate why measurements vary and processes that lead to variation in data
- Examine graphical representations of variability; use graphs to compare the variability of more than one data set.
- Focus on the bumps and clumps that appear in some graphs, and what they indicate about variability in the middle of a data set
- Promote awareness of both the overall spread and where the majority of the data are distributed.
- Examine measures of center and how measures of variability are based on spread from the center, recognizing how measures of variability are most informative in the context of a measure of center.
- Determine relative characteristics (e.g., resistance) of different measures of variability for different types of distributions, and when it makes sense to use particular measures as summaries of variability for particular distributions.

Bakker (2004) studied student reasoning about the shape of a distribution utilized two hypothetical learning trajectories in the lesson design. The first is described as:

> The overall goal of the growing samples activity as formulated in the hypothetical learning trajectory for this fourth lesson was to let students reason about shape in relation to sampling and distribution aspects in the context of weight. The idea

> was to start with students' own ideas and guide them toward more conventional notions and representation. (p. 68)

Bakker (2004) had students make up larger and larger samples and compare their predictions with actual data, as well at look at the shape in larger data sets with a continuous distribution.

Summary

The studies that have focused on student reasoning about variability in the areas of sampling, probability, distributions, and comparing distributions with a variety of levels of students and teachers have been reviewed in this chapter. The SOLO taxonomy provided a progression, or possibly a learning trajectory, that student reasoning might make as they learn to reason about variation.

Several studies have focused on introductory college students' reasoning when comparing distributions, however, these were typically small studies with a pretest-intervention-posttest design or studies that identified the types of misconceptions present in students' reasoning (delMas & Lui, 2005; Cooper & Shore, 2008; Meletiou & Lee, 2002; Inzunza, 2006; Slauson, 2008). Very few longitudinal studies have been completed (Watson, 2001; Watson & Kelly, 2004) and none of these longitudinal studies have focused on distributional reasoning of college students during an introductory statistics course, even though comparing distributions is one of the topics built up to in an introductory course. This study intends to address the lack of research on the change in students' reasoning about variability over the course of an introductory statistics course by answering the following open question in statistics education:

How does students' reasoning about variation when comparing distributions change as they progress through an introductory college-level statistics course?

Chapter 3

Methodology

The research question that guided this study was: how does students' reasoning about variation when comparing distributions change as they progress through an introductory college-level statistics course? This chapter will provide rationale for the design of the study, and describe participant selection, data collection, and data analysis procedures.

Design

To answer the research question, a mixed methods design was implemented. Creswell (2013) defined mixed methods research as:

> An approach to inquiry involving collecting both quantitative and qualitative data, integrating the two forms of data, and using distinct designs that may involve philosophical assumptions and theoretical frameworks. The core assumption of this form of inquiry is that the combination of qualitative and quantitative approaches provides a more complete understanding of a research problem than either approach alone. (Creswell, 2013, p. 4)

In considering the main research question, qualitative research can illuminate the spectrum of ways in which students learn to reason about variation, and quantitative research can potentially confirm which of these ways happen most often. The different ways of reasoning, as well as confirmatory evidence of these ways, can be found using qualitative and quantitative methods together. In addition, quantitative and qualitative data were analyzed simultaneously so that each type of data would inform the other.

The research question was posed with regard to a learning process. This necessitated that this study be longitudinal in order to study the changes over time in the reasoning of students. The study took place over the course of one semester as students took Introductory Statistics, and was observational.

Location of Study and Background

This study took place at a mid-sized doctoral-degree granting university in the Pacific Northwest. The University's focus is primarily liberal arts education, although there are several programs in the sciences and mathematics at the undergraduate and graduate level.

One of the three Introductory Statistics courses is taught by the Department of Mathematical Sciences (the other two are housed in the psychology and sociology departments). The course taught by the Mathematics Department consisted of three 50-minute lectures and one 50-minute lab section each week. Each semester, there were two lectures of about 250 students each with 16 lab sections of about 30 students. Both the lectures and all lab sections were highly coordinated. In Fall 2013, the lectures were taught by two instructors (including the researcher) who coordinated daily activities. The lecture material was contained in a course pack that students could print or purchase. The course pack contained examples, notes, and computing information for calculators and statistical software. The course pack was developed and refined over many years by statisticians from the department. Instructors would go through the course pack and add hand-written notes during each class period. The lab sections were taught by four graduate teaching assistants and one adjunct faculty member. Each week, the lab instructors and lecture instructors met to discuss the lab activities, homework and worksheet grading rubrics, and other topics. There were three exams spaced approximately evenly throughout the semester. The third exam was the final exam, but contained a minimal amount of material covered on the first two exams. Both lecturers used the i-clicker response system throughout the lectures in order to motivate student attendance and active learning in the lecture setting; however, the two instructors did not always ask the same questions, nor did they grade these questions in the same manner. Students also completed 12 worksheets in the lab sections, seven written 1-2 page homework assignments, and an online homework assignment for each chapter (a total of 18 chapters covered in the course). In the end, the majority of the students' grades were based on exams (69%), with homework, participation, and worksheets making up the rest of the grade (31%).

The course used *Intro Stats* by DeVeaux, Velleman, & Bock, 4th edition, along with the online homework system *MyStatLab* by Pearson. The main topics covered included: exploratory data analysis, linear regression, data collection, randomness and basic probability, central limit theorem, confidence intervals, and hypothesis testing for one and two proportions and means. The first exam covered exploratory data analysis and the second exam covered the remaining topics except for inference, which was covered on the third (final) exam.

Institutional Review Board

This study was approved by the Institutional Review Board at the University of Montana on March 1, 2013 (see Appendix A).

Pilot Study

A pilot study was completed by the researcher one semester before data collection. The goal of the pilot study was to make sure that interview and survey questions were asked clearly and that the amount of time students would spend answering the surveys and interviews was reasonable. Students were asked to complete one online survey and give permission to use their exam results. Of the students who gave their consent, 53 completed both the survey and exam. In addition, six of these 53 participants were interviewed one time. The changes made from the pilot study to the actual study survey included: simplification of the dot plot questions, inclusion of additional histogram questions, a reduction in the number of times students were asked to explain their answers, and re-ordering of the questions to go from simplest to most difficult. It was found in the pilot study that the context of the questions during the interviews greatly affected the depth of student answers. In order to minimize the extent of students' memory of particular questions in which they were highly interested, the contexts were varied throughout the surveys; however, the graphical displays stayed constant.

Participant Selection

All students who took Introduction to Statistics from the Mathematics Department at the University in the fall of 2013 were invited to participate in the study. A brief announcement describing the study was made during both lectures in the second week of class. Those students who were willing to participate in the study filled out an online consent form and demographic information through an online survey system outside of the online course system. On the consent form, students were also asked to indicate whether they would be willing to participate in a series of short interviews. After the period to fill out the online consent form passed (one week), fewer students than desired had chosen to participate in the study, so a paper consent form was also made available in the two lectures. The paper consent form lacked the question indicating interest in participating in the interviews since the interview students had already been chosen (see Appendix A). A total of 136 students filled out the combined online and paper consent forms.

Interview participant selection. Interview participants were selected from the students who indicated they were willing to be interviewed and had responded to the first online survey. All students in the course (regardless of their participation in this study) were instructed to complete the survey as part of their participation grade. Of the 43 students who were willing to be interviewed, 26 completed the first online survey. Using stratified purposeful sampling (Onwuegbuzie & Leech, 2007), 21 students with a large range of answers (from completely incorrect to completely correct) on the first Moodle assignment who also indicated that they were willing to be interviewed were recruited to participate in three interviews over the course of the semester. Purposeful sampling was used in the qualitative part of the research due to the constraint of small sample size (a very non-representative sample is possible with simple random sampling) and since random sampling was not feasible due to the necessity of a wide range of responses (Miles & Huberman, 1994), and also because these chosen individuals "can purposefully inform an understanding of the research problem and central phenomenon in the study" (Creswell, 2007, p. 125).

Of these 21 interview invitations, twelve students scheduled a first interview. Of these twelve scheduled students, one failed to attend the scheduled meeting. After rescheduling, this student failed to meet for the second time. One other student only participated in the first interview and stopped attending the course before the second interview, and did not respond to other interview requests. The remaining ten students completed all three interviews.

Data Collection

Data was collected from a series of three online surveys that were part of regular course assignments, a series of three tasked-based interviews with the purposeful sample. Course performance (grade) data and demographic information were also collected. The online surveys provided both quantitative and qualitative data, while the task-based interviews yielded only qualitative data. A diagram of the time sequence of data collection with observed sample sizes can be found in Figure 12.

Online survey data. During the course, students completed three online surveys through Moodle (the online course interface), which were comprised of multiple choice and open-ended questions. These questions focused on students' conceptual understanding of variation and asked students to reason about variation in histograms, dot plots, and bar graphs, as well as to explain their reasoning on these graphs (the first survey can be found in Appendix B). Each survey consisted of ten problems with one, two, or three individual questions related to each problem. Students were asked to judge the variability, "without making calculations", on all ten problems, and to explain their answers to three problems on each survey. Only three questions were chosen for qualitative answers in order to keep the survey to a reasonable length. Table 8 gives each question on the surveys with its various contexts. On the actual graphs for each question, the only alterations were the contextual labels; the scales were not altered.

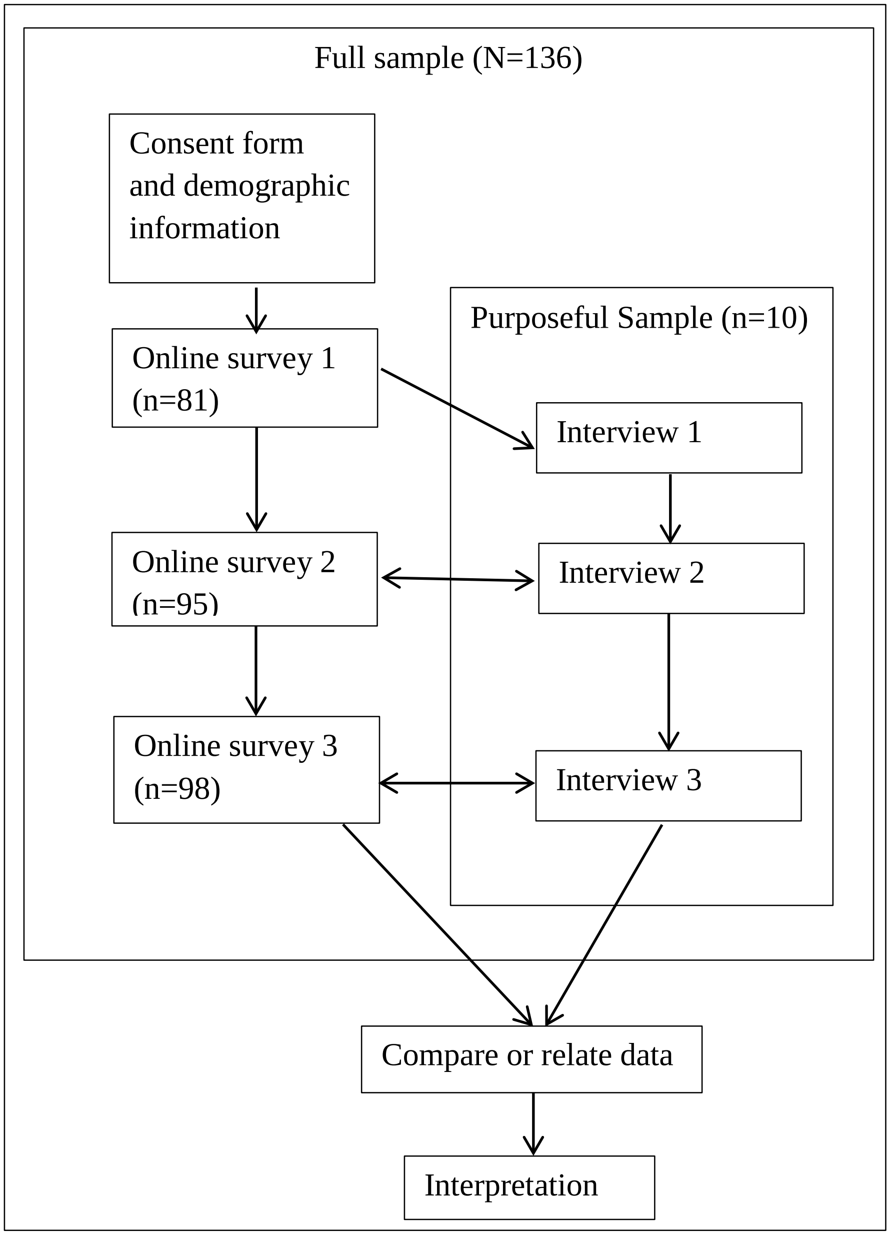


Figure 12. Data Collection Timeline. This figure indicates the order in which surveys and interviews occurred as well as the sample sizes and sample groups.

The first question, which included a bar graph about blood types of different ethnicities, and the last question involving a histogram were from Cooper and Shore (2010). Both questions' contexts were altered for the second and third surveys. The remaining questions were adapted from questions contained in Cooper and Shore (2010)

Table 8

Contexts of survey questions by type of graphs and relationship between graphs for each survey

Question type	**Survey 1**	**Survey 2**	**Survey 3**
Bar graph	Blood type by ethnicity	Type of pet by city	Favorite grocery store by city
Dot plot, different ranges	Number of pets	Times donated blood	Number of pets
Dot plot, translation	Number of children in a family	Number of exams during finals week	Number of children in a family
Dot plot, reflection	Number of bedrooms	Number of classes per semester	Number of bedrooms
Histogram, Bimodal and Z	Heights of students	Arm spans of students	Heights of students
Histogram, Uniform and T	Time to run 400 meters	Age of grandparents' death	Time to run 400 m
Histogram, Skewed and T	Hours spent going to class, studying and working	Distance from hometown to university	Exam scores
Histogram, Bimodal and Uniform	Basketball games scores	Lifespan of pet in months	Cell phone bill charge
Histogram, Skewed and Uniform	Money spent on groceries	Quiz grades	Money spent on groceries
Histogram, Z and T	Exam scores	Price of Halloween costume	Weekly minutes exercised

and delMas and Liu (2005). Some of these questions were reformulated into three dot plot questions in order to specifically allow students to have the individual data values, and five others were given as histograms with the clear lack of the original data values. In delMas & Liu (2005), students were led to believe that the data lay in the center of the bar instead of there being many possible data values for each data point within the bar. This would allow students to think of each histogram as a dot plot instead of not knowing enough information to make a dot plot without being given the data, and hence avoid reasoning about the presence of uncertainty.

Interview data. Ten students were interviewed three times over the course of the semester corresponding to each online survey, typically during the week before each of the exams. All interviews were audio recorded and later transcribed by the researcher. Interviews lasted between 6 minutes and 23 minutes and were conducted by the researcher. During each interview, students were asked to answer five of the online survey questions and explain their reasoning. Interview questions were chosen based on responses from the online survey. It was decided to only ask five questions in order to keep the interview to a reasonable length of time and to not frustrate students. The interviewer asked probing questions to student responses in order to better understand students' learning processes; however the interviewer attempted to not give feedback on the correctness of student answers until after all interviews were completed. A sample interview protocol is included in Appendix C.

Demographic data collection. In order to stream-line data collection, demographic information was collected immediately after students filled out the study consent form (online) or on the back of the consent form (paper). Data collected included: sex, major, class standing, approximate grade point average (G.P.A.), previous mathematics experience, socioeconomic status, and ethnicity. See Appendix D.

Course performance data. By giving consent to participate in this study, students gave permission for their course grades to be used. The data included: each exam score, online homework average, participation grade (included online surveys and daily i-clicker score), lab worksheet average, hand-in homework average, overall course average, and final letter grade. These data, along with the collected demographic information, made it possible to explore probable explanation of variation on students' reasoning and helped to provide information as to how representative the group of students who completed all three surveys was to the whole sample.

Validity. Construct validity is “the extent to which a test truly measures a proposed psychological ability or skill and is related to an underlying theory or model of behavior” (Salkind, 2009, p. 302). Construct validity was met through the pilot study, discussion of survey questions with a statistician, and the researcher’s own statistical knowledge. The pilot study provided feedback on how feasible and appropriate the survey questions were, which contributed to some adjustments to the questions as previously discussed. After developing the survey questions, the researcher met with a professor who is a statistician and had instructed the Introductory Statistics course many times. Together they went through the survey questions and revised them appropriately. In addition, the researcher has taken many graduate level statistics courses, passed a Ph.D. preliminary exam in statistics, and instructed as well as served as a teaching assistant in the Introductory Statistics course several times.

Data analysis

Due to the mixed methods design of this study, both qualitative and quantitative data were collected and analyzed.

Qualitative data. Interviews and survey explanation questions were coded using the SOLO model before being analyzed using descriptive statistics. The SOLO model was chosen based on its use in other studies about reasoning about variability (Peters, 2011; Reading, 2004; Turegun, 2011; Watson, 2009; Watson, Callingham, & Kelly, 2007; Watson, Collis, Callingham, & Mortiz, 1995) and the researcher’s view that it was the best possible model to describe the sophistication of students reasoning about the questions included in the survey. Inter-rater reliability was conducted with a second coder to increase reliability.

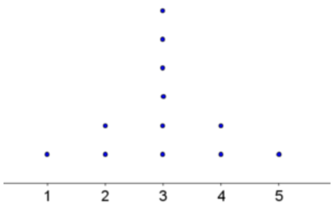
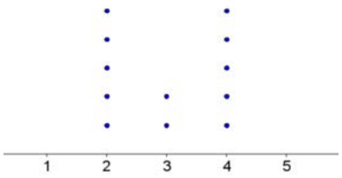
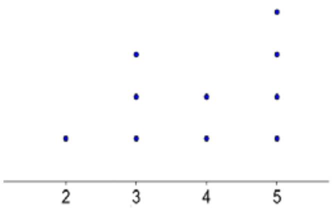
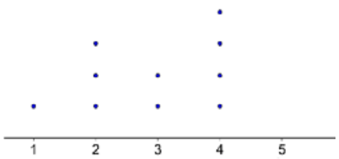
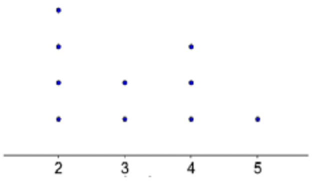
Quantitative data scoring. The survey questions fit into three categories: bar graphs, dot plots, and histograms. Only question 1 included bar graphs. For this question, students who chose the correct answer for the most variability and the least variability received 1 point. All other answers including partially correct answers received 0 points.

For the dot plot questions, Table 9 shows the dot plots from the questions and the appropriate measures of variability. On question 2, the first graph has a larger standard deviation and range, while the second graph has a larger IQR. While choosing either of these graphs as having more variability from these calculations demonstrates understanding of variability, based on the course content which encouraged students to use IQR instead of range and standard deviation when data were not mound-shaped, the best answer was to choose the graph with the larger IQR as being more variable and all other answers received 0 points. On questions 3 and 4, students who chose “no difference” received 1 point, and any other answer received 0 points since all of the measures of variability are equal.

In order to justify the correct answers on the histogram questions, data were created for each histogram in the following way: the smallest possible amount of variability, the largest possible amount of variability, and an approximately even distribution within each bar (as the qualitative data showed was commonly how participants’ dealt with a lack of data.) A summary of the histograms with these distributions is in Table 10. Note that the data were created to fit the histograms and to not overcomplicate the situation, each was rounded to the nearest tenth. There are an infinite number of other possible distributions for these histograms; however, these were the distributions used to decide which answers were best for the quantitative analysis.

Table 9

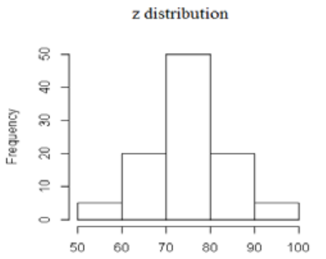
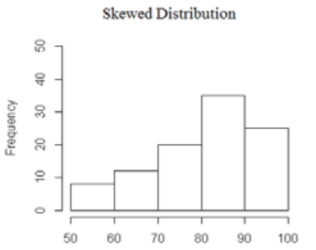
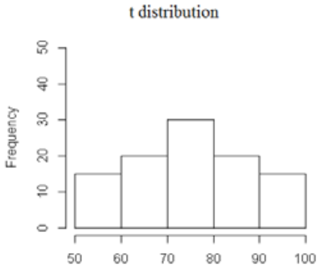
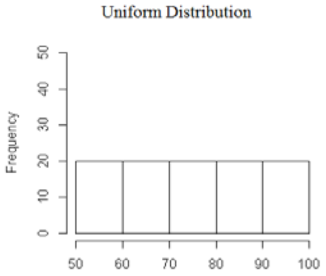
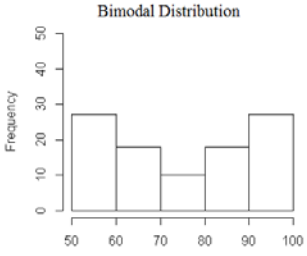
Dot Plots used in survey questions 2, 3, and 4 on all three surveys, and their measures of variability

Questions	Graph	Standard deviation	IQR	Range
2 (group 1)		1.04	1	4
2 (group 2)		0.95	2	2
3 (group 1) 4 (group 1)		1.10	2	3
3 (group 2)		1.10	2	3
4 (group 2)		1.10	2	3

From this information, the distribution with the least variability is the *z*-distribution, then the skewed distribution, then the *t*-distribution, then the uniform distribution, and lastly, the bimodal distribution is the most variable. This is assuming that only one of the sets of distributions (largest, even, and smallest) occurs, which is how the participants typically

Table 10

Largest possible, even, and smallest possible distributions of histograms used in survey questions 5 through 10 on all three surveys.

Survey Question	Graph	Measure	Largest Possible	Even distribution	Smallest Possible
5 (class 2) 10 (class 1)		Standard deviation	12.4	9.4	5.8
		IQR	15.0	10.0	5.0
		Range	49.9	48	30.1
7 (class 1) 9 (family 1)		Standard deviation	16.5	12.6	8.4
		IQR	22.4	17.5	2.6
		Range	49.9	48.6	30.1
6 (group 2) 7 (class 2) 10 (class 2)		Standard deviation	16.7	13.0	8.9
		IQR	29.9	19.8	10.1
		Range	49.9	49.4	30.1
6 (group 1) 8 (team 2) 9 (family 2)		Standard deviation	18.5	14.5	10.0
		IQR	29.9	24.8	10.1
		Range	49.9	49.5	30.1
5 (class 1) 8 (team 1)		Standard deviation	20.5	16.2	11.5
		IQR	49.9	31.5	30.1
		Range	49.9	49.4	30.1

considered the situation, as found in the qualitative data. These decisions make the correct answer to question 5 to be class 1, question 6 to be group 2, question 7 to be class 2, question 8 to be team 1, question 9 to be family 2, and question 10 to be class 2. Students with each of these answers received one point.

Data were first analyzed by looking at each question separately and looking for a difference between the whole group of students who took any of the surveys and the group of students who took all three surveys. Total student scores for each survey were also analyzed comparing the entire group with those who took all three surveys.

Triangulation. Triangulation, which is evidence from multiple sources (Creswell, 2013), was achieved through the use of three surveys and three interviews. The same ten sets of graphs were used for all three surveys, with differences in the context of the questions. In the interviews, five questions out of the ten from the survey were chosen to be discussed with the participants. Table 11 indicates which questions were used in the surveys and interviews for qualitative responses. Participant responses to all of these qualitative questions were used to find themes and appropriate SOLO model placement.

Table 11

Qualitative data collection of qualitative responses on surveys and interviews by question number

Question	Surveys	Interviews
1	1	2
2	1, 3	1, 2, 3
3		
4	2	
5	3	1, 2, 3
6		1, 2, 3
7		1, 3
8		
9	2	1
10	1, 2, 3	2, 3

Summary

The focus of this study was to answer the research question: how does students' reasoning about variation when comparing distributions change as they progress through an introductory college-level statistics course? This chapter explained the design of this study, as well as participant selection, data collection, and data analysis procedures. In the next chapter, the results of data analysis are presented.

Chapter 4

Results

Throughout this chapter, quantitative and qualitative results will be interwoven to answer the research question: how does students' reasoning about variation when comparing distributions change as they progress through an introductory college-level statistics course? First, the demographics of the sample are presented, then the coding process of the qualitative data and inter-rater reliability, results on the progress made through each of the four types of survey questions and the progress of the interviewees are presented, followed by the overall progress made on the survey.

Demographics

In order to better understand the group of students participating in the study, participants were asked to report demographic information when giving consent. The questions can be found in Appendix D. Of the 136 students who gave consent to participate in this study, three chose not to report any demographic information. This gave a typical sample size of 133 for the whole group for the following responses to demographic information, however the sample size is 136 for participants' final grades since they did give consent to the use of their final grades in the course.

Participants were either in the lecture section of the researcher or one other instructor. Of the 133 participants, 31% were in the researcher's lecture, and 35% of the students who completed all three surveys were in the researcher's lecture section. Of the ten participants completing all three interviews, 50% were in the researcher's lecture. The rest of the participants were in the other lecture.

The majors of the participants can be found in Table 12. In each group of participants (whole group, completed three surveys, and interviewees), the percentages in Business, Health, Humanities, Science, Technology, Engineering, and Mathematics (STEM), and other majors were similar. The highest percentage of participants were majoring in STEM fields, then Business, and then Health. Very few participants in this study had majors in the humanities or other areas.

Table 12

General categories of majors of participants in study as reported by participants

	Business	Health	Humanities	STEM	Other or undeclared	Number of participants
Whole group	29%	20%	5%	38%	8%	133
3 surveys	33%	19%	3%	41%	3%	63
Interviewed	30%	30%	0%	40%	0%	10

The introductory statistics course that all 133 participants were in is considered a sophomore-level course, and the largest percentage of all participants were sophomores (39%). There was also a high percentage of Juniors (32%) and Seniors (17%) in the whole group, although there were also Freshmen (7%) and Post-baccalaureate students (5%). The ten interview participants were mostly higher-ranked students with two post-baccalaureate and four seniors, but the group interviewed also included a freshman, two sophomores, and a junior.

Students' prior mathematics knowledge was captured through information on the previous mathematics courses they had taken, both at the university and in high school. At the university, approximately half of the students had taken a finite math course, half had taken a pre-calculus course, and half had taken a calculus course, as seen in Table 13.

Table 13

Participants' previous college mathematics courses as reported by participants

	Intermediate Algebra	Finite and Linear Mathematics	College Algebra, College Trigonometry, or Pre-calculus	Applied Calculus or Calculus	Introduction to Statistics	Number of participants
Whole group	21%	48%	44%	45%	3%	133
3 surveys	17%	44%	50%	52%	2%	63
interviewed	0%	50%	40%	50%	0%	10

A higher percentage of female students chose to participate in this study than male students (62% female to 36% male, 2% chose not to report their sex out of 133 participants). This was not consistent with the interview students (4 were female, 5 were male, and one chose not to report a specific sex).

The university this study was conducted at is known for having a high number of non-traditional students. In the whole group of 133 participants, 32% of the students were under 20 years old, 56% were between 20 and 29 years old, and 12% were 30 or older. In the group of the ten interviewed participants, 3 were under 20 years old, 3 were between 20 and 29 years old, and 4 were 30 or older.

Due to lack of variation in ethnicity at a university where the vast majority of students are Caucasian, ethnicity was not reported in order to preserve the anonymity of the participants.

The whole group of participants typically had high G.P.A.s (75% above 3.0) but 25% had G.P.A.s under that, or were without a G.P.A. (See Table 14). The G.P.A.'s of the interviewees were noticeably higher; all participants had G.P.A.s above a 3.0, and 7 had a G.P.A. above 3.5. Note that this data was self-reported during data collection.

Table 14

Participants' grade point averages prior to study (self-reported)

G.P.A.	3.50 to 4.00	3.00 to 3.49	2.50 to 2.99	2.49 and under	No G.P.A.	Number of Participants
All participants	44%	31%	19%	4%	2%	133
3 surveys completed	49%	30%	18%	3%	0%	63
Interviewees	70%	30%	0%	0%	0%	10

The final course grades of the participants were relatively high, with 49% of participants with an A or B (includes "+" and "-" grades), and 8 of the 10 interview participants getting an A or B. See Table 15.

Table 15

Final course grades in introductory statistics course of participants

	A	B	C	D	F	No grade	Number of Participants
All participants	20%	29%	35%	5%	7%	4%	136
3 surveys completed	28%	28%	33%	8%	2%	2%	64
Interviewees	40%	40%	10%	10%	0%	0%	10

Qualitative Data Coding

Since the researcher conducted and transcribed the interviews (in addition to completing a pilot study) many student responses shaped the way she looked at the quantitative survey data. These experiences, along with the qualitative survey answers collected in this study and prior studies utilizing the SOLO taxonomy, led to the creation of a rubric (Table 16) for understanding responses on the survey and interview tasks. This framework was used because the SOLO levels matched student reasoning on the survey and interview tasks and was a meaningful way to describe the complexity of the responses. Due to the nature of the questions, the cycles were in the concrete symbolic mode and included pre-structural (in the ikonic mode), uni-structural, multi-structural, relational, and extended abstract (in the formal mode). After transcribing each interview, survey responses and interview responses were coded according to the SOLO framework.

Table 16

SOLO Framework with appropriate task responses for the ten survey questions in this study

Mode	Level	“more variability”	“same variability”	“less variability”
Ikonic	Pre-structural	- Heights of bars differ more (Code H) - Larger mean - Has a peak (Code P) - Larger distribution - Range on y-values (Code R)	- Symmetric (Code S)	- More consistent
Concrete symbolic	Uni-structural	- More categories (Code C) - Larger spread, range (Code R) - More different numbers per category - More results different from mean - Bimodal (Code S)	- Same number of responses in each category (Code C) - Same spread, range (Code R)	- Fewer categories (Code C)
	Multi-structural	- Heights of bars differ less (Code H) - More evenly spread out (Code V) - More [equal numbers] in each bin (Code V) - Doesn’t have a peak (Code P)		- Heights of bars differ more (Code H) - Less evenly spread out (Code V)
	Relational	- More in tails or further from center/mean/median (Code T) - Larger IQR (Code IQR) - Larger standard deviation, variance (Code SD)	- Mirror images, opposites (Code M) - difference to mean equal/ equal mean absolute deviations	- Less in tails (Code T) - More close to center/mean/median - Smaller IQR (Code IQR) - Smaller standard deviation (Code SD)
Formal	Extended Abstract	- Not able to decide based on given information because relational answers conflict		

Pre-structural responses. Students responding at the pre-structural level recognized that there was variation in the displays but were unable to describe it in a way that was statistically meaningful. These students often thought that when the heights of the bars were more variable, there was more variability in the display; however, the opposite it actually true. When prodded during interviews, these students may have showed a lack of understanding of histograms. For example, the tenth question on the first survey (see Figure 13) asked "Which class has more variability in their exam scores?" and also required students to explain their answers. A pre-structural response as given by Student 6 was "Class 1" with the explanation: "class 1 goes up and down much more than class 2. We can see this in the graph". This response is pre-structural because it describes there being more variability when the bars in the histogram are more variable, which is not the case.

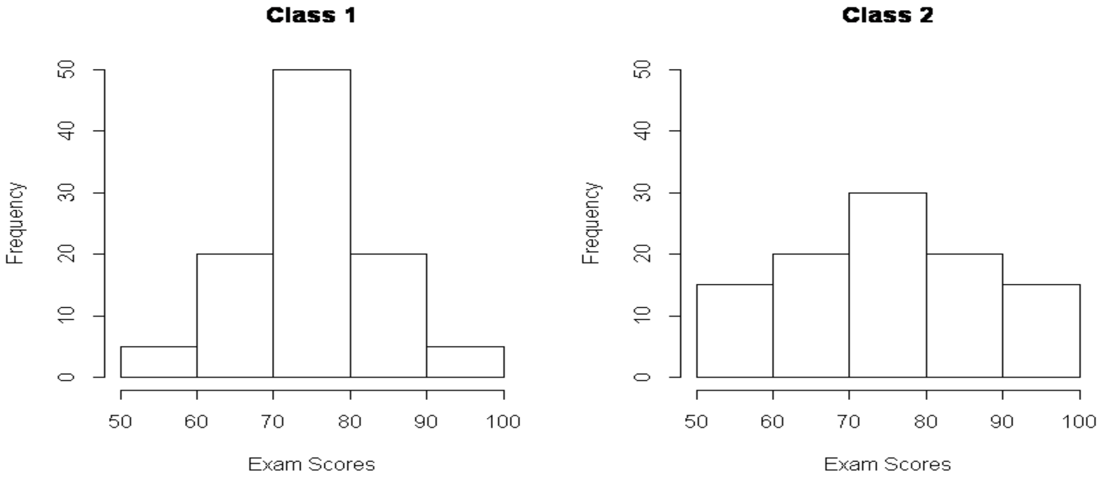


Figure 13. Question 10 from Survey 1

Uni-structural responses. Students responding at the uni-structural level recognized variation in one dimension, such as a bigger range or when there are a greater number of categories. Since range only refers to one dimension (horizontal) and can therefore be greatly affected by outliers, this choice of measurement is uni-structural.

Students responded that a dot plot with a larger range was more variable regardless of the placement of the dots. For example, also referring to the problem from Figure 13, a uni-structural response from Student 98 was “No difference” with the explanation “same range, same variability”. Although it is possible (and likely) that the two graphs have different ranges when calculated from the original data, students were not provided the original data from these graphs, and so the estimate that they have the same range is appropriate.

Multi-structural responses. Students responding at the multi-structural level could incorporate two or more features of variation such as in the horizontal and vertical direction, but were unable to put these ideas together. They responded to questions where the distributions had the same range by explaining that the graphs with bars that had similar heights had more variability, while graphs with larger differences in heights had less variability; however, these students were unable to see a difference if the bars in a histogram were reordered. For example, also referring to the problem from Figure 13, a multi-structural response from Student 21 was “class 2 has more variability because every category is at least 10%, while class 1 has two categories (50-60 and 90-100) with less than 10%”. This response was denoted multi-structural because this student pointed out that the bars of the histogram were more even in height, which caused greater variability; however, that the position of the bars on the x-axis is important was not referred to. The positioning could very easily change the measures of spread and change which graph would have more variability.

Relational responses. Students responding at the relational level were able to make connections between the center of the data and the spread of the data. This was either by discussing the proportion of data in the tails compared to the proportion of data in the center or by using the IQR or standard deviation. Both the IQR and standard deviation are measures of variability that take into account both center and spread of the distribution. However, students at this level were not able to recognize that one distribution could have a larger IQR, while the other distribution could have a larger standard deviation such as in Question 2 (see Appendix B). In addition, students at this level were also unable to notice that, without the actual data, histograms such as those in Question 10 (see Figure 13) could possibly either have the larger IQR or larger standard deviation. For example, a relational response was given by Student 106: “The IQR for class 2 would be larger than for class 1, so class 2 has more variability”. The student noticed that there was a larger range for the middle 50% of data in class 2, however, without the actual data, the IQR could possibly be larger for either graph. This lack of observation is the difference between a relational response and an extended abstract response.

Extended abstract responses. Students responding at the extended abstract level were able to recognize when the IQR and standard deviation were in conflict (such as in Question 2). They were also able to recognize that the histograms in Questions 6 through 10 did not necessarily lead to one or the other having a higher standard deviation or IQR. Either graph can have a smaller range, IQR, or standard deviation when extreme (but possible) examples are created. No students gave an extended abstract response for the question in Figure 13.

A summary of these levels and the modes in which they fall (consistent with Biggs and Collis (1991)) can be found in Table 16. Within each level, there were three categories, “more variability”, “same variability”, and “less variability.” Two of the questions (3 and 4) had the same measures of spread, and Question 1 asked for the most and the least variable distributions. In addition, some students chose the most variable distribution after describing the features of the least variable distribution.

Inter-rater reliability. The researcher extensively trained a second rater on the coding process outlined in Table 16. Then both individuals independently coded ten student responses from each one of the six unique questions that had qualitative data over all three surveys. A random sample of size ten was drawn from each set of survey question answers. In addition, both raters independently coded three full interviews. Both raters gave both the SOLO level and an answer type code (H, P, C, R, S, M, V, T, IQR, SD) when one was appropriate. On the six questions that survey answers were given, the two raters agreed on the SOLO levels 90% of the time (54 out of 60), and the answer type codes 86.7% of the time (52 out of 60). All disagreements were resolved except for one where raters discovered that in the particular answer, it was impossible to know whether the student made a calculation error or a conceptual error.

For the interview data, a random sample of three interviews was chosen, with each interview containing five questions, to be coded by both raters. On the questions where interview responses were given, the two raters agreed on the SOLO levels 84.2% of the time (16 out of 19), and the answer type codes 84.2% of the time (16 out of 19). All disagreements were discussed between the two raters and resolved.

Progress by Question Type on Survey

There were four types of questions on the surveys based on the type of graphs in the questions: bar graphs, dot plots, histograms with a uniform distribution, and histograms without a uniform distribution. Each of these question types is discussed through both quantitative and qualitative data, including the survey responses, the qualitative responses, and the progress made on the type of question.

Bar graphs. Only the first question on all three surveys (see Appendix B) contained bar graphs. In this question, students were asked to choose the bar graph with the most variability and least variability, effectively ranking the three graphs. Participants were also asked to explain their responses to this question during the first survey and the second interview. Although the question content was changed with each survey, the third graph (Graph C) always had the most variability due to the nearly even amount of data in each of the four categories and the first graph (Graph A) always had the least variability due to nearly all of the data being in a single category.

Survey responses. The bar graph questions had a very high correct response rate. Table 17 shows the responses to the most variability question from the 64 students who answered all three surveys. The percentage of students choosing the correct answer (Graph C) decreased over time, however, overall the percentage of students with the correct answer was high (ranged from 72% to 84%).

Table 17

Student responses to the graph with the most variability for the bar graph question (Question 1) for the n=64 students completing all three surveys

Survey	A	B	C (correct)	I'm not sure	No Difference
1	9%	3%	83%	3%	2%
2	11%	2%	84%	0%	3%
3	13%	8%	72%	2%	6%

The same pattern was also seen in Table 18, for the responses to the least variability question. These percentages of correct responses (Graph A) were actually higher than the responses for the most variability question and showed less of a decrease in performance over time. In addition, none of the participants chose graph B.

Table 18

Student responses to the graph with the least variability for the bar graph question (Question 1) for the n=64 students completing all three surveys

Survey	A (correct)	B	C	I'm not sure	No Difference
1	86%	0%	11%	2%	2%
2	84%	0%	14%	2%	0%
3	80%	0%	14%	2%	5%

Qualitative responses. The qualitative responses from survey 1 were coded using the SOLO model (see Table 19). Due to the nature of the bar graph question, it was not possible to give a response above the multi-structural level without much explanation and some other examples, which did not occur when students were asked to answer a short survey.

Table 19

SOLO levels of the qualitative responses to the bar graph question (Question 1) from survey 1 for all participants

SOLO Level	Number of participants
Pre-structural	10%
Uni-structural	11%
Multi-structural	75%
No explanation	4%
Number of responses	81

Eight students gave pre-structural responses. For example, Participant 67 found the range of the bars in each graph, and the bar graph with the largest of these values was considered the most variable, which was not a statistically valid way to respond to the question:

> Question 1: most variability means have most range in their blood types. And the range is equal to max - min, the max is around 100%, while min only 2%
>
> Question 2: least variability means most data are same or the range is short, the max is 40%, and min is around 10%. (Participant 67)

Nine students gave uni-structural responses, which were often based on having the same number of bars in each graph. For example, Participant 68 stated: “Each graph has the same variable options O, A, B, AB and each have at least one person for each variable, so therefore the variability is the same for each ethnic group” (Participant 68). While this response did show some understanding of variability, by only mentioning the presence of each category it did not take into account all of the information given in the graphs.

Most participants responded at the multi-structural level, and the majority of these students responded by explaining that the graph with more variability was more evenly spread out or had more [equal numbers] in each bin. For example, Participant 11 responded that graphs that are more evenly spread out have more variability, while a graph that has mostly one category has very little variability:

> The Japanese graph's data is most evenly spread over the four different blood types which means there is the most diversity among the people as far as what kind of blood they have. The Mayan graph indicates that almost everyone has O blood and the other types are almost nonexistent. (Participant 11)

These responses show the different ways that participants reasoned about bar graphs, both using correct and incorrect ideas.

Bar graph question progress. The overall progress of the participants completing all three surveys on the two multi-choice bar graph questions (identifying the graph with the most variability and the least variability) can be seen in Table 20. From the first to the third survey, the number of correct responses decreased, however, this was not a statistically significant decrease (p = 0.0507) under a paired t-test comparing the difference in means from the first survey to the third survey (mean difference of −0.17189, and standard error of 0.10340).

Table 20

Number of students with correct answers to the two bar graph questions on each survey for students completing all three surveys

Number of Points	Survey 1	Survey 2	Survey 3
0	9	10	13
1	2	0	5
2	53	54	46

Note: The maximum score was 2 points

Table 21 shows the progress of the individual students classified as “Improved”, “No change”, or “Declined”. Note that 77% of the students made no overall change when answering the bar graph questions over the course of three surveys.

Dot Plot Questions. There were three dot plot questions on each survey (see Appendix B for these questions). Question 2 had students compare a dot plot that had a larger range (graph 1) with one that had a larger IQR (graph 2). Question 3 had students compare the same dot plots, with an overall translation by 1 unit. Question 4 had students compare a dot plot with its mirror image.

Survey responses. On question 2, most students chose the incorrect graph with the larger range over the graph with the larger IQR as having more variability (see Table 22).

Table 21

Overall progress and survey-to-survey progress of students completing all three surveys (n=64) on the two bar graph questions

Overall Progress	Number	First to Second Survey	Second to Third Survey	Number
Improvement	4	Improved	Improved	0
		Improved	No change	4
		No change	Improved	0
		Declined	Improved	0
		Improved	Declined	0
No change	49	No change	No change	45
		Declined	Improved	2
		Improved	Declined	2
Overall decline	11	Declined	Improved	0
		Improved	Declined	0
		No change	Declined	8
		Declined	No change	3
		Declined	Declined	0

Table 22

Student responses to the graph with the most variability on the dot plots with the larger range and the larger IQR (Question 2) for the students completing all three surveys (n=64)

Survey	A (larger range)	B (larger IQR, correct)	I'm not sure	No Difference
1	57	5	1	1
2	50	4	7	3
3	52	4	6	2

On question 3, the vast majority of students recognized that the translated graphs had the same variability, although there were a few (3 on survey 1, 6 on survey 2, and 3 on survey 3) that chose one graph or the other graph, or who were unsure. See Table 23.

On question 4, a vast majority of students still recognized that there was no difference in the variability in the mirror image graphs, however, there were more students choosing the skewed left graph, the skewed right graph, or who were unsure (7

Table 23

Student responses to the graph with the most variability on the dot plots with the larger mean and the smaller mean (Question 3) for the students completing all three surveys (n=64)

Survey	A (larger mean)	B (smaller mean)	I'm not sure	No Difference (correct)
1	2	1	0	61
2	2	4	0	58
3	2	0	1	61

on survey 1, 9 on survey 2, 3 on survey 3) than had an incorrect response in question 3. See Table 24.

Table 24

Student responses to the graph with the most variability for the skewed left and skewed right dot plots (Question 4) for the students completing all three surveys (n=64)

Survey	A (skewed left)	B (skewed right)	I'm not sure	No Difference (correct)
1	2	1	4	58
2	5	2	2	55
3	0	1	2	61

Qualitative responses. Over the three surveys, students explained their answers to a dot plot question three times. The responses were coded by SOLO level and a summary can be found in Table 25.

Table 25

SOLO levels of the qualitative responses to the dot plot questions (Questions 2 and 4) from surveys for all participants

SOLO Level	Question 2, Survey 1	Question 4, Survey 2	Question 2, Survey 3
Pre-structural	4%	16%	11%
Uni-structural	82%	40%	76%
Multi-structural	9%	1%	4%
Relational	0%	39%	1%
No explanation	5%	4%	8%
Total	81	95	98

When explaining their answer to question 2, most students responded at the uni-structural level in both the first and third surveys, although about 11% gave pre-structural responses during the third survey. In the uni-structural responses, 46 were coded “range” and 14 were coded “categories.” For example, a uni-structural response with “range” focused on making a decision for group 1 as having more variability due to having a larger range: “The responses range from 1 to 5 giving a higher spread of responses” (Participant 40). The code of “categories” focused on choosing group 1 as having more variability due to the number of categories with data: “In group 1, each category has at least one dot, which group 2 only has dots for 2, 3, and 4 pets” (Participant 21). Pre-structural responses often focused on irrelevant ideas such as “The mean number of dots is the same” (Participant 37), or had incorrect ideas such as having a greater difference in the heights of the bars: “One group of six in group one versus two groups of five in group two makes for a total greater variability in group one” (Participant 65). Another participant with a pre-structural response incorrectly thought that “All have the same range” (Participant 98) perhaps due to both graphs having the same x-axis even though graph 2 did not have data points in each part.

When explaining their answers to question 4, students were split between the uni-structural (40%) and relational levels (39%), although there were 15 participants that gave pre-structural responses. A uni-structural response to this question typically mentioned that both graphs had the same range: “Both groups have the same range so even though group 1 is skewed to the left and group 2 is skewed to the right their variation is the same” (Participant 42). A relational response often described the graphs as being mirror images of each other, thus having the same amount of variability: “They both have the same number of responses spread in the same pattern (just backwards) over the same range” (Participant 11).

Overall progress on questions with dot plots. On the three questions using dot plots, there was not a large change in the number of correct responses from the group of students completing all three surveys, as seen in Table 26. Most students answered two out of three questions correctly, and these were generally questions 3 and 4, which had equal amounts of variability.

Table 26

Number of correct responses to the three dot plot questions on the three surveys for participants taking all three surveys (n=64)

Points	Survey 1	Survey 2	Survey 3
0	1	4	2
1	5	6	2
2	55	51	56
3	3	3	4

For the overall progress made by students, the majority had no overall change (51 out of 64 participants), see Table 27. From the first survey to the third survey on the dot plot questions, overall, 51 students had the same number correct, 7 students had a higher number correct on the third survey than the first, and 6 students had more correct on the first survey than the third. In addition, 41 students stayed consistent on all three surveys with no change from survey to survey.

Histograms with uniform distributions. There were three questions on each survey containing a uniform distribution along with another distribution (see Appendix B for these questions). Question 6 compared the uniform distribution with a *t*-distribution. Question 8 compared the uniform distribution with a bimodal distribution. Question 9

Table 27.

Overall progress and survey-to-survey progress of students completing all three surveys on the three dot plot questions

Overall Progress	Number	First to Second Survey	Second to Third Survey	Number
Improvement	7	Improved	Improved	1
		Improved	No change	2
		No change	Improved	3
		Declined	Improved	1
		Improved	Declined	0
No change	51	No change	No change	42
		Declined	Improved	6
		Improved	Declined	3
Overall decline	6	Declined	Improved	2
		Improved	Declined	1
		No change	Declined	1
		Declined	No change	2
		Declined	Declined	0

compared a uniform distribution with a skewed distribution. One of the common misconceptions that arose on these questions was that students would base their decision on the variability of the heights of the bars. This caused them to believe that a uniform distribution had no variability due to the lack of difference in the heights of the bars instead of taking the position and height of the bars into account.

Survey responses. Over the three surveys, at least half of the students thought that the *t*-distribution had more variability than the uniform distribution in Question 6 (see Table 28). Only about a quarter of the students correctly thought that the uniform distribution was actually more variable than the *t*-distribution.

On question 8, students compared the bimodal distribution and the uniform distribution. More than half of the students thought the bimodal distribution was more variable; however, this question never asked students to explain their reasoning. Based on the other two questions, students often based their responses on the graph being more

Table 28

Student responses to the graph with the most variability between the uniform and t-distributions (Question 6) for the students completing all three surveys (n=64)

Survey	Uniform (correct)	*t*-distribution	I'm not sure	No Difference
1	18	33	1	12
2	15	39	5	5
3	14	32	8	10

spread out, or having a greater difference in the heights of the bars. In this situation, both of these responses would lead them to choosing the bimodal distribution and so without any qualitative responses it is unclear whether the reason they chose the bimodal distribution was correct or not. See Table 29.

Table 29

Student responses to the graph with the most variability between the bimodal and uniform distributions (Question 8) for the students completing all three surveys (n=64)

Survey	Bimodal (correct)	Uniform	I'm not sure	No Difference
1	42	12	2	8
2	42	11	4	7
3	33	14	7	10

Question 9 asked students to compare the skewed distribution and the uniform distribution. Again, over half of the students thought the skewed distribution was more variable than the uniform distribution, and only around a quarter of the students thought the uniform distribution was more variable than the skewed distribution. See Table 30.

Table 30

Student responses to the graph with the most variability between the skewed and uniform distributions (Question 9) for the students completing all three surveys (n=64)

Survey	Skewed	Uniform (correct)	I'm not sure	No Difference
1	42	14	2	6
2	39	15	4	6
3	33	16	5	10

Qualitative responses. The whole group only explained their answers to one question involving a uniform distribution, which occurred on question 9 on survey 2. The SOLO levels of these responses can be found in Table 31.

Table 31

SOLO levels of the qualitative responses to the histogram questions containing a uniform distribution (Question 9) from survey 2 from all participants

SOLO Level	Question 9, Survey 2
Pre-structural	57%
Uni-structural	9%
Multi-structural	18%
Relational	6%
No explanation	9%
Total	95

Pre-structural responses often described the uniform distribution as having no variability “because there is no variability in class 2” (Participant 10), or described the skewed distribution as having more different frequencies in the bars and therefore, more variability: “Class 1 has more variability because the frequency is not constant like in Group 2” (Participant 29). Most of the uni-structural responses pointed out that the range of both distributions was the same, so the variability in both graphs was the same. The multi-structural responses explained that the uniform distribution was more variable due to having a more even distribution: “There is a more even distribution” (Participant 41). A relational response also discussed the overall position of the skewed distribution: “they have an equal amount of spread out numbers, but class 2 is more equal in the grades, where class 1 is more towards the right” (Participant 35).

The interviewees were asked to explain their responses to question 6 during each interview. Over the course of three interviews, there were participants responding at each of the SOLO levels. Emily gave a pre-structural response focusing on the same frequencies in the uniform distribution as being less variable:

> Well, there's 20 in each age group, they're like the same chance, well, in this model, having a grandparent die in 50 to 60, and 90 to 100, so I think there would be less variability in that versus here where all the proportions aren't the same, the frequencies aren't. (Emily, Appendix E)

Tim gave a uni-structural response focusing on both graphs having the same range "Oh, right, the variability is the same, because they're all dying between 50 and 100. So the variability between 50 and 100 is the same as the variability between 50 and 100" (Tim, Appendix E). Josh explained that the uniform distribution was more variable than the *t*-distribution at the multi-structural level by describing a more even distribution as being more variable: "I think this one has more variability just because each group is equally represented…" Brian gave a relational response by taking into account the data near the center and the data away from the center: "This would be interesting to see everyone has the same amount of running, although I feel like this one would have more variability, because it's completely distributed across, versus a lot more in middle versus the sides" (Brian, Appendix E). Mark arrived at an extended abstract response through a discussion with the interviewer in which he was able to see that the measures of variability such as the IQR were not necessarily fixed and were unknown due to not having the actual data:

> Mark: I believe I went with group 1 again the first time I did it, and it was because there is an equal number of people in every group, so, your middle 50% would be, 20, 40, 60, most of these, these two are here, the middle 50% is going to be this, half of each of these, yeah, half of each of those.

Interviewer: So do you know where those data points are in those bars? Like you're kind of telling me half of the bar, right? Do you know that people are evenly spread out in that bar?

Mark: No, you don't because it's a histogram and it's the people between 60 and 70 and so, I guess that does make a difference in a way, but I don't think it makes a big enough difference that it would change my answer though, I'm still going to go with group 1.

Interviewer: So if you were making up the numbers of these running times, for group 1 and group 2, could you make up the numbers so that group 2 would have a larger IQR than group 1 and still have the pictures lined up being the same?

Mark: Yeah, if you put them all at 60 or at 90, or 61 and 89, you could make the IQR larger, so, yeah.

Overall progress on questions with uniform histograms. On these three questions, over half of the participants completing all three surveys got only one question correct, and over a quarter got no questions correct on the third survey. (See Table 32).

Table 32

Number of correct responses to the three histogram questions containing a uniform distribution on the three surveys for participants taking all three surveys (n=64)

Points	Survey 1	Survey 2	Survey 3
0	9	10	16
1	38	39	34
2	15	12	13
3	2	3	1

From the first survey to the third survey on the histogram questions including a uniform distribution, overall, 45 students had the same number correct, 5 students had a higher number correct on the third survey than the first, and 14 students had more correct on the first survey than the third. In addition, over half of the students had no change from the first to the second and the second to the third surveys. See Table 33.

Table 33

Overall progress and survey-to-survey progress of students completing all three surveys (n=64) on the three histogram questions containing a uniform distribution

Overall Progress	Number	First to Second Survey	Second to Third Survey	Number
Improvement	5	Improved	Improved	0
		Improved	No change	1
		No change	Improved	3
		Declined	Improved	1
		Improved	Declined	0
No change	45	No change	No change	36
		Declined	Improved	3
		Improved	Declined	6
Overall decline	14	Declined	Improved	2
		Improved	Declined	0
		No change	Declined	9
		Declined	No change	2
		Declined	Declined	1

Histograms without Uniform Distributions. There were three questions on each survey that asked students to compare histograms that were not uniform (see Appendix B for these questions). On question 5, students compared a bimodal distribution and an (approximate) *z*-distribution. On question 7, students compared a skewed distribution and an (approximate) *t*-distribution. On question 10, students compared an (approximate) *z*-distribution and an (approximate) *t*-distribution.

Survey responses. Question 5, comparing the variability of the bimodal and *z*-distributions, was the question with the most extreme cases of histograms with different variability. It was not possible to create a data set that would create the same histograms presented in the problem where the IQR of the bimodal distribution would be smaller than those measures with the *z*-distribution (see Table 11). Over the three surveys, there was a small improvement, with nearly half of the students choosing the correct answer (bimodal) during the third survey. See Table 34.

Table 34

Student responses to the graph with the most variability between the bimodal and z-distribution (Question 5) for the students completing all three surveys (n=64)

Survey	Bimodal (correct)	*z*-distribution	I'm not sure	No Difference
1	28	22	2	12
2	29	19	2	14
3	31	17	7	9

Question 7, comparing the skewed distribution and the *t*-distribution was more difficult, with only about a quarter of the students choosing the correct answer (*t*-distribution) over the three surveys. See Table 35.

Table 35

Student responses to the graph with the most variability between the skewed and t-distribution (Question 7) for the students completing all three surveys (n=64)

Survey	skewed	*t*-distribution (correct)	I'm not sure	No Difference
1	30	18	8	8
2	29	16	9	10
3	27	16	3	18

Question 10 asked students to compare the *z*-distribution with a *t*-distribution. Only around 40% of the students chose the correct response (*t*-distribution). See Table 36.

Table 36

Student responses to the graph with the most variability between the z-distribution and t-distribution (Question 10) for the students completing all three surveys (n=64)

Survey	*z*-distribution	*t*-distribution (correct)	I'm not sure	No Difference
1	22	28	5	9
2	30	25	2	7
3	23	25	6	10

Qualitative responses. Participants explained their answers to Question 10 on each survey, as well as to Question 5 on the third survey. In Table 37, it was seen that each question had the most pre-structural responses, but each level (except extended abstract) was present during each survey. It is interesting to note that about twice as many students responded at the uni-structural level on question 5 as compared to question 10.

Table 37

SOLO levels of the qualitative responses to the histogram questions not containing a uniform distribution (Questions 5 and 10) from all three surveys from all participants

SOLO Level	Question 10, Survey 1	Question 10, Survey 2	Question 5, Survey 3	Question 10, Survey 3
Pre-structural	38%	38%	31%	40%
Uni-structural	9%	9%	22%	10%
Multi-structural	27%	28%	25%	26%
Relational	15%	11%	12%	11%
No explanation	11%	14%	10%	13%
Total	81	95	98	98

Question 5 responses were present at each SOLO level. Pre-structural responses tended to again focus on the variability of the bar heights: “There is more variability in the heights of class 2 because even though it is distributed across the same heights as class 1, there is greater differences in frequency within Class 2” (Participant 72). Uni-structural responses tended to describe the graphs having the same range and therefore the same variability, or chose group 1 based on it being bimodal: “Group 1 is more variable because it is bimodal” (Participant 88). Multi-structural responses either used the heights of the bars being less variable as having more variability: “Class 1 has more variability because the frequency of each height are closer together than in class 2” (Participant 91) or found the bimodal distribution to be more evenly spread out: “there is a more even distribution in class 1” (Participant 14). Relational responses also took into account the location of the bars: “there are more towards the edges and less in the center” (Participant 81) or estimated a larger measure of variability: “Class 1 has a bigger IQR” (Participant 82).

Participants also responded to Question 10 (*z*-distribution and *t*-distribution) at all levels in the SOLO model. Pre-structural responses included misinterpreting the heights of the bars and miscalculating range: “Greatest range” (Participant 62) or using this same idea of bar heights without a calculation: “Class 1 data goes up and down much more than class 2” (Participant 6). Uni-structural responses often referred to both graphs having the same range or spread: “Spread of the distributions is the same” (Participant 108). Multi-structural responses included the *t*-distribution as being more spread out: “Group 2 has more variability because the frequency is more spread out” (Participant 21) or the *z*-distribution having nearly all of the data in one place: “The majority of group #1 are between 70-80, therefore there is less variability” (Participant 38). Relational responses focused on a larger amount of data further from the mean in the *t*-distribution: “More people farther from the mean” (Participant 2) or the *t*-distribution having a larger measure of variability “The IQR for class 2 would be larger than for Class 1, so Class 2 has more variability” (Participant 106).

Overall progress on questions without uniform histograms. Table 38 shows the progress made on the histogram questions that did not contain a uniform distribution. From the first survey to the third, more participants got all questions incorrect, but there was also an increase in the number of students getting two or three of these three questions correct.

Table 38

Number of correct responses to the three histogram questions not containing a uniform distribution on the three surveys for participants taking all three surveys (n=64)

Points	Survey 1	Survey 2	Survey 3
0	25	27	30
1	16	14	7
2	11	13	16
3	12	10	11

From the first survey to the third survey on the histogram questions not including a uniform distribution, overall, 26 students had the same number correct, 20 students had a higher number correct on the third survey than the first, and 18 students had more correct on the first survey than the third. See Table 39. These three questions overall had more movement (both positive and negative) than the bar graph, dot plot, and uniform distribution questions.

Table 39

Overall progress and survey-to-survey progress of students completing all three surveys (n=64) on the three histogram questions not containing a uniform distribution

Overall Progress	Number	First to Second Survey	Second to Third Survey	Number
Improvement	20	Improved	Improved	2
		Improved	No change	4
		No change	Improved	10
		Declined	Improved	2
		Improved	Declined	2
No change	26	No change	No change	16
		Declined	Improved	4
		Improved	Declined	6
Overall decline	18	Declined	Improved	3
		Improved	Declined	2
		No change	Declined	5
		Declined	No change	6
		Declined	Declined	2

In addition, question 10 had a qualitative response on each survey. Overall, 4 participants increased, 29 students maintained, and 9 students decreased from survey 1 to survey 3 in the SOLO level of the responses (see Table 40.) Note that 22 of the 64 students responding to this question on all three surveys did not explain their answer at least once and were not included.

Table 40

Overall progress and survey-to-survey progress of students completing all three surveys on qualitative responses comparing the z-distribution and t-distribution (Question 10) as measured by the SOLO model (n=42)

Overall Progress	Number	First to Second Survey	Second to Third Survey	Number
Improvement	4	Improved	Improved	0
		Improved	No change	2
		No change	Improved	2
		Declined	Improved	0
		Improved	Declined	0
No change	29	No change	No change	18
		Declined	Improved	3
		Improved	Declined	6
Overall decline	9	Declined	Improved	2
		Improved	Declined	0
		No change	Declined	2
		Declined	No change	5
		Declined	Declined	2

Progress of Interviewees

The progress made by the ten interviewees fell into four themes: students who improved their reasoning about variability (*n*=2), students who maintained their reasoning about variability (*n*=3), students who decreased in their reasoning about variability (*n*=2), and students who were inconsistent in their reasoning about variability (*n*=3).

Improvement. Two interview participants, Josh and Megan, overall improved their reasoning about variability over the course of three interviews.

During Josh's first interview, he gave responses from the pre-structural level to the relational level. In his second interview, Josh gave responses from the pre-structural level to the multi-structural level. During his third interview, Josh gave responses mainly from the multi-structural level and one response from the uni-structural level and none from the pre-structural level. Josh consistently responded to the second question (larger range and larger IQR dot plot question) at the uni-structural level by basing his response on the range or the number of categories. On question 7 (skewed distribution and *t*-distribution), he went from a pre-structural response in the first interview: "These values are all different, they vary so much more…" indicating that he was looking at the heights of the bars as being more variable to a multi-structural response in the third interview where he indicated that the graph that was closer to uniform was more variable. Josh also improved from interview 2 to interview 3 on question 10 (*z*-distribution and *t*-distribution). During interview 2, he again based his response on variability meaning the heights of the bars were more variable, and then in interview 3, he chose the correct graph and explained his reasoning: "I'd say this would be more variable because it's going toward the flatter model." Overall, Josh showed improvement by no longer responding to questions with pre-structural responses during the third interview.

Megan had quite advanced responses, even in her first interview. She responded to questions at the uni-structural level up to the extended abstract level. During her second interview, her lowest responses were at the multi-structural level and again went up to the extended abstract level. By her third interview, all responses were at the relational and extended abstract levels. She arrived at the extended abstract level on the last question that was asked during the interview. The improvements made over the course of the interviews occurred on questions 2, 5, and 6.

In question 2, the larger range versus larger IQR dot plot question, Megan gave a uni-structural response based on the graph's range during the first interview: "Because it is spread out over a further period even though it has a lot more in the middle, I would go with group 1" (Megan, Appendix E). Then during the second interview, her response was at the relational level: "So it would be more spread out, away from the mean, I think" (Megan, Appendix E). She maintained this level of response during the third interview, even though she changed her answer back to the first graph as being more spread out and then focused on the standard deviation, which was larger for graph 1 (but not by a great amount):

> That's what makes me wonder about group 2, that there's more that aren't right at the mean. Like this one has a lot right at the mean and not many far away, but they're spread out overall. So that's what kind of trips me up because I don't know which one would have a bigger standard deviation without calculating it. (Megan, Appendix E)

On question 5 (bimodal distribution and *z*-distribution), Megan also improved her reasoning from a multi-structural response in the first interview to a relational response in the second and third interviews. In the first interview, Megan described the graph having more variability as being "more spread out" (Megan, Appendix E) which was a multi-structural response, taking into account the heights of the bars and the location of the bars. In the third interview, Megan described the graph having more variability as having “more away from the mean” (Megan, Appendix E) which now was taking into account the position of the bars from the center and was at the relational level.

On question 6 (uniform and *t*-distributions), Megan went from a multi-structural response in the first interview to a relational response that became an extended abstract response in the second and third interviews when questioned further by the interviewer. In the first interview, she chose the uniform distribution as having more variability because “the bars were the same” (Megan, Appendix E). In the second and third interviews, she again chose the uniform distribution as having more variability than the *t*-distribution, but explained that if the bin width was changed to a smaller size, that might change which had more variability: “Yeah, you probably could because if these little bars were a different interval, it might change how they looked and that would make me change my mind” (Megan, Appendix E). This foresight in how she did not have the data to be able to see how this might be different and changed her overall decision on these graphs is at the extended abstract level of reasoning.

Both Josh and Megan showed an improved reasoning about variability from the first interview to the third interview.

Maintained. Three participants, Brian, Peter, and Allison, overall maintained their knowledge and reasoning about variability over the course of the three interviews. Brian continued to have results in the uni-structural, multi-structural, and relational levels over all three interviews. Peter had nearly all pre-structural responses during all three interviews. Allison tended to have pre-structural and uni-structural responses and then through discussion, could often reason at a multi-structural or relational level, but then was unsuccessful in translating that improvement to other questions during those interviews.

Brian stayed consistent with his answer to question 2 during all three interviews, basing his answer on the range of the two graphs and identifying the dot plot with the larger range as having more variability, which was a uni-structural response. On the questions about histograms, he gave a response at the multi-structural level, almost always using the description of a graph being more evenly distributed as having more variability. During each interview, on at least one of the questions about histograms, Brian also had a response in the relational level. For example, on question 6 (uniform and *t*-distributions) in the first interview, he chose group 1 as having more variability due to: "…it's completely distributed across versus a lot more in the middle versus the sides" (Brian, Appendix E). In the second and third interviews, Brian had a relational response to question 10 (*z*-distribution and *t*-distribution), comparing the distributions to determine which had more in the middle: "Because a lot more people are in group 1 are hogging up space in the middle, like 70 to 80 dollars on costumes, versus there's a more even distribution of how money is spent in group 2" (Brian, Appendix E). Overall, Brian's responses stayed rather consistent during each interview.

Peter's responses during all three interviews also stayed consistent, although at the pre-structural level. Overall, Peter tended to compare the variability of heights of the bars instead of what the bars represented. Peter was able to show an understanding of the bars in the first and second interviews: "… so you have 20 people running this in 50 to 60, and 20 in 60 to 70…" However, he felt like there was not much variability in these since it was always 20 people in each bar, a pre-structural response. Peter did not always explain what the bars represented in the graphs correctly. During the second interview, he thought the mean of the money spent by group 2 in question 10 was $25 (when all of the totals were between $50 and $100) because he was looking at the mean of the bar heights. During the course of three interviews, Peter consistently reasoned at the pre-structural level.

Allison overall maintained her inconsistency within each interview and was unable to take reasoning from one question and translate it to another. During her first interview on question 5, she reasoned that the *z*-distribution and the bimodal distribution had the same amount of variability because: “They all range from 55 to 80 inches.” Then during interview 2, she started by repeating that the graphs had the same variability before changing her answer to the bimodal distribution due to “They’re all kind of in the middle, the majority of them in the middle” which was a relational answer taking the center and spread into account. Again, during the third interview, she went from thinking they had the same amount of variability to choosing the *z*-distribution as having more to finally end with the bimodal distribution as having the most variability due to similar reasoning as in interview 2. Allison did stay consistent with responding to question 2 (larger range versus larger IQR dot plots) with a uni-structural response each time, and responding to question 6 (uniform and *t*-distributions) with a pre-structural response.

Overall, Brian, Peter, and Allison were consistent in their reasoning during all three interviews.

Decreased. Two participants, Tim and Hannah, had overall decreases in their knowledge and reasoning about variability over the course of three interviews.

During interview 1, Tim gave multiple uni-structural, multi-structural, and relational responses. During interview 2, Tim ended each question with a uni-structural response, and by interview 3, he gave a few uni-structural responses and mostly pre-structural responses. For example, on question 5 (bimodal distribution and *z*-distribution), during the first interview, Tim chose Class 1 as having more variability because, "It's just more numbers spread out over a greater range" (Tim, Appendix E) and then explained that, "This one seems to have a bunch of frequencies in the middle, and then it tapers off and this one has a lot of frequencies in the ends, and then it kind of drops down in the middle" (Tim, Appendix E) which was a relational level response. In interview 2, he started by choosing the group with greater variability in the bars (a pre-structural response), then moved to describing that the other graph has more variability due to the "scattering" (a multi-structural response), before ending with thinking the graphs have the same amount of variability due to having the same range (a uni-structural response). In the third interview, Tim started off by stating that the variability was the same due to the same range (a uni-structural response) but ended with deciding that the class with more variability in the bar heights was more variable (a pre-structural response).

A more direct decrease in responses was seen in Tim's responses for question 6 (uniform distribution and *t*-distribution). In the first interview, he gave a relational response. In the second interview, he gave a uni-structural response. In the third interview, he gave a pre-structural response. Tim did give consistent responses to question 2 (larger range and larger IQR dot plots), each time basing his response on the range of the bar graph. However, overall, the quality of his answers decreased from quite good to incorrect over the course of the three interviews.

Hannah also showed an overall decrease in knowledge and reasoning, however it was not as drastic as Tim's. Even from the first interview, Hannah's responses were pre-structural, uni-structural, and multi-structural, however, this decreased to only pre-structural and uni-structural responses by the third interview.

This decrease can be illustrated by Hannah's responses to question 6 (uniform and *t*-distribution). In the first interview, Hannah was very intrigued by the context of the runners, being a runner herself. She chose the uniform distribution as having more variability over the *t*-distribution, a multi-structural response. During the second interview, the context of question 6 was age of grandparents' death, she chose the *t*-distribution as having more variability due to the heights of the bars, which was a pre-structural response. However, she did realize that appropriate calculations would be the IQR and standard deviation. During the third interview, this question had the same context as the first interview (runners) but did not have the same effect on her reasoning as the first time: Hannah chose the *t*-distribution as being more variable without a clear explanation (pre-structural response).

Another decrease for Hannah was in question 2 (larger range versus larger IQR dot plots). During the first interview, she used range to make her decision (uni-structural), in the second interview, she based her decision on the number of categories with data (also uni-structural), and during the last interview, she ended up comparing the difference in the heights of the bars in an incorrect way (pre-structural response): "I guess this one has more in this area and this one doesn't."

Hannah and the researcher had a long conversation during the second interview on question 5 (bimodal and *z*-distribution), where she was able to get to a multi-structural response, she referred to class 2, “Then there would be a lot in the middle.” Additionally:

> Class 1 would have been like a lot of short people and a lot of tall people and very few in the middle. So then when you look at it that way, you’re just like “oh, then its class 1.”

However, she was not able to transfer this reasoning to the two subsequent questions she was asked about, and this reasoning also did not come up in the third interview.

Both Tim and Hannah had significant decreases in their reasoning during the second and third interviews.

Inconsistent. Emily, Mark, and Nicole were inconsistent over the three interviews. This was present through a large decline in their descriptions from the first to the second interview, but then improved back to where they were at in the first interview during the third interview.

During Emily’s first interview, her responses were at the pre-structural, uni-structural, multi-structural, and relational levels. During the second interview, Emily’s responses were between the pre-structural and multi-structural levels, but not at the relational level. During the third interview, her responses again went from the pre-structural level up to the relational level.

Emily’s responses to question 5 (bimodal and *z*-distribution) moved around from interview to interview and also within the interviews. During interview 1, she gave a multi-structural response basing her answer on the bimodal graph being more uniform than the *z*-distribution. In interview 2, she first decided that the bimodal graph had more variability due to it being more uniform (a multi-structural response) before switching this reasoning to the uniformness causing less variability (a pre-structural response): “If they’re kind of around the same proportion, then there is less variability versus this one. I change my answer” (Emily, Appendix E). Finally, during interview 3, Emily started by saying that the *z*-distribution graph had more variability due to the frequencies having more variability (pre-structural level) but then changed her response to the bimodal distribution “Because there’s more on the ends, further away from the center, the 65-70 inch range and you have fewer in the center range” (Emily, Appendix E) which was a relational response.

Emily also had inconsistent answers on question 6 (multi-structural and pre-structural), question 7 (multi-structural and pre-structural), and question 10 (pre-structural, multi-structural and relational). In addition, during each interview on questions involving histograms, her responses included a pre-structural response and a multi-structural response, showing that her reasoning on these questions was not consistent.

Mark also was inconsistent in his reasoning during the three interviews. During Mark’s first interview, his responses were all at the relational and extended abstract levels. During interview 2, his responses were at the uni-structural, multi-structural, and relational levels. During interview 3, Mark went back to having all his responses at the relational level.

In each interview, Mark tended to respond with only one or two different types of reasoning. In interview 1, that reasoning was based on the IQR. In interview 2, Mark based his reasoning on the range and also more data being away from the mean. In interview 3, he based his reasoning on the standard deviation and having more data away from the mean.

During Nicole's first interview, she responded at the uni-structural, multi-structural, and relational levels. During her second interview, she continued to respond at the uni-structural and multi-structural levels but did not get to the relational level. In her third interview, she responded at the multi-structural and relational levels. Her reasoning on these questions was not as consistent within each interview as Mark's responses. For example, in interview 2, Nicole's response to question 5 (bimodal and *z*-distributions) was at the multi-structural level, focusing on how evenly spread out the distribution was:

> I'd say class 1 has more variability because in each of the categories, each of the different ranges, or boxes, there are more evenly spread out between all of them than in class 2 where most of the people have arm spans between 65 and 70 inches.

During that same interview on question 6 (uniform and *t*-distribution), she responded: "I'd say class 1 has more variability because all of the bars are the same height unlike in class 2" focusing her reasoning on the heights of the bars, which was a different type of response also at the multi-structural level.

Emily, Mark, and Nicole were all inconsistent over the three interviews, often declining from interview 1 to interview 2 and then improving from interview 2 to interview 3 in their reasoning.

Overall Survey Progress

There were 64 students who completed all three surveys, which were scored overall on a correct/incorrect basis. The first question (bar graph question) asked students to identify the graphs with both the greatest and least variability, so both of these questions were included, which gives a total point allowance of 11 points. Table 41 shows that the mean student scores decreased over the surveys and the standard deviation increased, however, this decrease was not statistically significant (p = 0.0946) when comparing the means of the first and third survey in a paired t-test (mean difference from the first to third survey was −0.3438 and had a standard error of 0.2590).

Table 41

Means and standard deviations of overall scores of students completing all three surveys (n=64) out of 11 possible points

	Survey 1	Survey 2	Survey 3
Mean	5.94	5.73	5.59
Standard deviation	2.10	2.20	2.28

Table 42 showed the progress made by each student from survey to survey and overall. Nearly half of the students showed an overall decline (30 of 64 students) on the entire survey score, which shows a decline in their reasoning about variability from the beginning of the course to the end of the course. However, nearly one third of students showed an overall improvement (21 of 64 students). While fewer students had no change between the surveys than when broken into the smaller categories of question type, this was expected due to having more questions overall. Table 42 showed that while there were some individuals that improved, overall, there was no improvement made by the participants as a whole, and possibly an overall decrease in their reasoning about variability during an introductory statistics course.

Summary

This chapter gave the results from the data from several different views. Participants' demographic information was described. The SOLO model was described in relation to the survey questions and inter-rater reliability results were given. Survey and interview questions were presented in four different groups based on the type of graph in the question including bar graphs, dot plots, uniform distribution histograms, and non-uniform distribution histograms. Interviewees were found to be in four different progression themes including: improved, maintained, decreased, and inconsistent. Finally, the overall scores on the survey showed a lack of significant improvement in the students' responses over the course.

Table 42

Overall progress and survey-to-survey progress of students completing all three surveys (n=64) on the entire survey

Overall Progress	Number of students	First to Second Survey	Second to Third Survey	Number of students
Overall Improvement	21	Improved	Improved	4
		Improved	No change	8
		No change	Improved	5
		Declined	Improved	3
		Improved	Declined	1
No change	13	No change	No change	5
		Declined	Improved	6
		Improved	Declined	2
Overall decline	30	Declined	Improved	6
		Improved	Declined	8
		No change	Declined	8
		Declined	No change	2
		Declined	Declined	6

Chapter 5

Discussion and Conclusions

The goal of this study was to examine the changes in students' reasoning about variation when comparing distributions as they progress through an introductory college-level statistics course. As was seen from the results of analysis in the previous chapter, there was an overall lack of improvement in students' reasoning about variability over the course of an introductory course, although there was not a statistically significant decline in their reasoning. This chapter describes the significance of this study, discusses the analysis from both the quantitative and qualitative data, compares this study with the literature, provides recommendations for college-level statistics instructors, acknowledges the limitations of this study, and provides recommendations for future research.

Study Significance

This study looked at reasoning about variability in a longitudinal manner with college-level students and helps to fill many gaps within the field of statistics education. In the existing literature, there are few studies with college-level students in statistics education, and there is a lack of studies in the area of reasoning about variation (Shaughnessy, 2007). This study also gave some examples of students' high-level responses in the SOLO model, which were not present in any of the previous studies (e.g., Reading & Shaughnessy, 2004). In addition, while most previous studies with college-level students in this area have been at most cross-sectional or measured students' reasoning before and after a teaching intervention, this study was longitudinal. The only longitudinal studies were done by Watson and Kelly (2004), who focused on elementary and secondary students, and Zieffler and Garfield (2009), who focused on changes in bivariate reasoning.

Discussion

There are a number of possible explanations for the overall lack of improvement on students' reasoning about variability. The structure of the particular course the participants were enrolled in is a likely factor, as is the dual role of the researcher as the teacher. In addition, students' tendency to turn abstract thinking into concrete thinking, as well as their misunderstandings when briefly glancing at histograms, often keep them from achieving higher-level reasoning. This lack of improvement also indicates the complexity and counter-intuitiveness of learning about variability in a distribution, which appears to require direct attention, which was not given to participants in this study.

Several aspects of the course that all participants took changed from the beginning to the end of the study, including the conceptuality of the content being taught, use of different measurers of variability, and the actual topics being discussed. While instructors attempted to keep the course conceptual, focusing on explanations rather than calculations, by the end of the course, when learning about hypothesis testing, it was clear to the instructors that students found hypothesis testing to be more focused on following the steps rather than understanding the concepts of the process. This may have contributed to more difficulties to reason conceptually in the surveys and interviews.

In addition to the decreasing of the conceptual nature of the course, the use of different measures of variability also changed. Although the course began by using different measures of variability, including range, IQR, and standard deviation, by the end of the course the only important measures of variability were the standard deviation and standard error. This may have caused students not to use other measures of variation after the first survey. Furthermore, students first completed the survey around the time that variability was being discussed in lecture, perhaps giving students the idea that the surveys would be on material currently being covered in class. When the course moved on to other topics but the second and third surveys essentially contained the same topics as the first survey, students may have been trying to relate the survey to the content being covered at the time, thus hurting their ability to answer the questions. Students were never told that the second and third surveys were essentially the same as the first one, even though the topics on the survey were not independently discussed again specifically during the lecture.

Another possible explanation for the lack of progress made by the students was that the researcher was also the teacher of a significant number of participants in the study. In the other lecture section of the course, the researcher also delivered the lecture six times when the other instructor had a medical problem. Overall, all students in the study may have viewed the researcher as an instructor of statistics content. In most courses, students receive straightforward feedback (e.g., correct or incorrect on their homework problems) from teachers on their knowledge, and this course was no exception for many assignments; however students were not given this kind of feedback on the survey or during the interviews from the researcher-teacher (although after the last interview, the researcher did give participants feedback if desired). This lack of feedback could have resulted in situations such as a participant assuming he/she did not correctly answer questions and looking for a different method, or assuming he/she was correct and not trying to reason in a more sophisticated way in the future. In addition, during interviews, if students reasoned at a low level, the researcher typically tried to ask them to explain different parts of the questions, and to continue to think about the question. However, when students thought in a higher level of reasoning, once they had finished their explanation, the researcher did not ask them to explain why a lower level was incorrect. If interviewees viewed the researcher as the teacher, this may have lead them to think that even without feedback, when the researcher allowed them to stop explaining their answer to a particular question, they must have arrived at the correct answer.

Students' responses were often quite concrete, showing that they viewed histograms as having an even distribution within each bar. No students responded at the extended abstract level in the survey responses, and only arrived at this level through continued discussion with the researcher during interviews. This lack of abstract reasoning is not unsurprising; even upper division mathematics majors often reduce abstract reasoning to concrete thinking (Abdalbaki, 2012). The participants in this study not only had less mathematical background than those in Abdalbaki's study (2012) but would likely have the same struggles in reasoning abstracting in statistics.

One of the common misconceptions participants had was looking at the variability in the bar heights instead of looking at the positions of the bars on the x-axis. This misconception occurred for Emily before she took the time to think about what the bars in the histogram represented. When she did this, she was then able to change her response from being at the pre-structural level to a relational level response, although this did not carry over to the following questions in the interview. This lack of immediate understanding of what the histogram was representing with regards to the data, which may have occurred in many of the participants when taking the survey, may explain the large number of incorrect or pre-structural responses in the histogram questions. Perhaps an additional question before the variability question where students are required to think deeply about what is being represented in the histogram would be beneficial to their responses about the variability. In addition, this example demonstrated the influence of the researcher on the interview participants.

Qualitative Data. The interviews showed that essentially all of the possible paths in reasoning about variability existed from the first to the third surveys: improved, maintained, decreased, and inconsistent. In addition, there was more than one student in each of these paths. Although the interviewees were not a random sample of the whole class, this result indicated that each of these paths occurred at least with some frequency. All of the surveys did take place after the topic was covered in class, so there is no indication as to how much the class coverage benefited the participants. In addition, it seems possible that the lack of feedback from the researcher-teacher during the surveys may have caused some students (such as Tim, who made a steep decline in his reasoning about variability over the three surveys) to doubt their reasoning and try to find other answers, even though interviewees were told that the researcher would not be giving them feedback on their responses in terms of the correctness. In the end, it appears that without additional specific instruction to students, they will not progress in their reasoning about variability, and often, even when figuring out one situation, the ideas did not transfer to future questions during that interview or in future interviews.

While most of the different demographic attributes of participants were spread rather evenly between groups of improvement type made by participants (improved, maintained, decreased, inconsistent), there was one characteristic that may have contributed to the progress made by students. Of the interviewees, the two students who improved both took pre-calculus (or college algebra or college trigonometry) and calculus (calculus I or applied calculus) while the other students had not taken both of these courses (see Appendix D for categories). A similar trend was also apparent in the students completing all three surveys; 67% of those who improved overall had taken a calculus course, 62% of those that stayed the same had taken a calculus course, and only 37% of those who declined overall had taken a calculus course. Calculus courses often contain abstract thought, and so this may have helped the students who had taken a calculus course in their reasoning about variability in this study.

Quantitative Data. The quantitative data was broken into four categories due to the type of graphs within the questions. On the bar graph questions, students typically answered quite well with at least 72% of the students taking all three surveys getting both questions correct on each survey. Responses were also reasonably correct on the dot plot questions, with at least 84% of the students taking all three surveys getting at least two of the three questions correct on each survey. This number of correct responses was particularly notable when taking into account that many of the responses for question 2 were based on the graph with the larger range as being more variable than the graph with the larger IQR. Due to the shape of the graphs, using the IQR to decide which had more variability was a better answer, and hence if students chose the graph with the larger range, it was marked incorrect, although they were at least using one of the measures of variability in their responses. The histogram questions were more difficult for the students. On the histogram questions containing a uniform distribution, more than 73% of the students taking all three surveys got none or only one of three questions correct on each survey. On the histogram questions without a uniform distribution, more than 58% of the students taking all three surveys got none or only one of three questions correct on each survey. Overall, it was possible that students could use more instruction on the variability present in histograms throughout an introductory statistics course. Perhaps this could be addressed by referring to standard deviation and its connection to variability during hypothesis testing, which would not only make hypothesis testing better understood, but also help students to better understand the variability in histograms. In addition, uniform distributions are not always addressed in an introductory statistics course, and this study makes it clear that they are not generally understood by students. Additional instruction and examples of uniform distributions may be helpful in increasing students' understanding of variability.

Comparison with Existing Literature. This study found results similar to those in previous studies both in the ways students described variability in distributions and the lack of progress made by students.

Many studies have already highlighted the misconceptions or reasoning students make when describing distributions. This study confirmed the presence of student misconceptions such as: that bell-shaped histograms with more variable bars have greater variability (Cooper & Shore, 2008; delMas et al., 2007; Meletiou & Lee, 2002); or that any two histograms with the same range have the same amount of variability regardless of other features of the distribution (Cooper & Shore, 2008). Many community college instructors also based their decisions about variability on the range of the distributions (Dabos, 2011).

The SOLO model used in this study also contains the same ideas as the Hierarchy of Reasoning about Variation that was used in a sampling distribution setting (Reading & Shaughnessy, 2004) although in their study the responses were broken up into fewer categories.

The lack of overall progress by students has been seen in other longitudinal studies in similar courses and with similar topics. Zieffler and Garfield (2009) completed a longitudinal study in a semester-long college introductory statistics course for social science students to understand how students' reasoning about bivariate data, specifically in a regression situation, changed throughout the course. Four classes of approximately 30 students each participated in the study and took the same survey four times. Two instructional sequences were used, with bivariate data either appearing after sampling and exploratory data analysis and before sampling distributions, probability, and statistical inference, or the same sequence with bivariate data at the end of the course. The largest increase in scores happened between the beginning of the course and before starting either distributions or bivariate data. The researchers attributed this to a general increase in statistical reasoning (Zieffler & Garfield, 2009).

Another possible explanation as to why students may not have made progress may be the plateau that Watson & Kelly (2005) found in performance of the Grade 9 students in their ability to explain variation in distributional situations. These researchers found increased ability in the younger students (Grades 3, 5, and 7) but found that there was no improvement in Grade 9 students. Noll and Shaughnessy (2012) also found a lack of a significant difference between successful middle school and successful high school students when reasoning about variability in sampling distributions. Although their study focused on sampling distributions, this may indicate that students at the college-level have already made the possible improvements in their reasoning about variability and explain the lack of progress found in this study.

In contrast to the results in this study, studies with younger students and with longer time frames than this study have found significant improvement in students' reasoning about variability when comparing distributions both longitudinally and through cognitive conflict. Watson (2001) found large improvements in reasoning about comparing distributions made by students three or four years after their initial participation while in grades 3, 5, 7, and 9. Watson and Kelly (2004) also found an improvement in 9th grade students' ability to reason about variability in probability situations two to three years later. The time periods in these studies were much longer than in this study, which may contribute to the increase in reasoning. In addition, most of the students in the study were younger than the students in this study.

An additional explanation for the lack of improvement in this study may be the lack of cognitive conflict, an important factor in other studies for quick improvement in reasoning. Watson (2002) tested the influence of cognitive conflict in a group of students reasoning about variability when comparing distributions. By watching videos of other students reasoning in a different way about the distribution, 57% of the students increased their reasoning quickly. Bakker (2004) also found large improvements in students' ability to reason about the shape of a distribution after a whole class discussion and individual questioning by interviewers on the topic. In this study, cognitive conflict was not present for students whose responses were multi-structural or above during interviews, and was not present at all on surveys. In addition, due to the lack of direct feedback from the researcher-teacher, students did not know how correct their responses were. Furthermore, while the topics of variability and distributions were presented in class, there was no discussion about different ways to think about these ideas. These aspects of this study may have contributed to the lack of progress in students' reasoning when compared with the studies of Watson (2001, 2002) and Bakker (2004).

Recommendations. Based on the results of this study, there are several recommendations for instructors of college-level statistics courses in order to enhance students' reasoning about variability in distributions. First, it would be beneficial to spend more time discussing histograms and the meanings of their different aspects, perhaps in an interactive activity either using a computer simulation or a classroom activity. Also, due to student success in the understanding of variability in bar graphs and dot plots, relating these situations with histograms so the reasoning process can then be expanded as the topics become more difficult. In addition, it may be helpful to use class discussion or cognitive conflict with students to get them to understand the possible views and different ways to reason about variability. Furthermore, it is important for students to see extreme situations of data that fit into particular histograms and see how these change the measures of variability. Lastly, students tend to have difficulty understanding standard deviation (delMas & Lui, 2005), so additional instructional materials and situations may be useful to enhance their understanding of this measure of variability and then hopefully transfer these concepts to their reasoning. With these changes, students may have a greater success with reasoning about variability in distributions after a college level introductory statistics course.

Limitations

Due to this study having both quantitative and qualitative components, there were many limitations. Neither the whole group of participants nor the interviewees were a random sample, in fact, the interviewees were chosen to try to represent the differences in students. While approximately 500 students were asked to participate in the study, only 136 gave their consent. Then, many students in this group of participants did not complete all three surveys, resulting in a somewhat small and biased sample of students willing to participate and complete three assignments that were a small part of their participation grade. This biased group of participants also occurred with the interviewees: of the 21 students contacted, only 12 students actually scheduled a first interview and only ten completed all three interviews. Only some of the students indicated that they were willing to be interviewed, and since this interview was conducted by someone they saw as a teacher, this likely had an influence on who was willing to participate. Fortunately, this group of ten interviewees were still rather diverse in their responses.

Future Research Suggestions

From the results of this study, it appears that the concepts covered in an introductory statistics course do not enhance students' reasoning about variability, so it may be necessary to use cognitive conflict or some other pedagogy to increase learning and study the effect of such interventions. It is also possible that the lack of improvement of students' reasoning may be attributed to design flaws in the surveys and interviews.

This study looked at a particular type of students, often with majors in STEM, business, or health sciences. There are certainly other groups of students who would benefit from deeply understanding variability and how to reason about it, such as pre-service and in-service K-12 teachers, who will be instructing students before college in statistics. Getting these teachers to understand variability at a high level would hopefully help them to teach students about this concept before they reached the college level. In addition, typically those pre-service teachers focused on elementary certification will not even take an introductory statistics course at the college level, but will be teaching variability in distributions beginning in the sixth grade (Common Core State Standards, 2010). In order to reach this population, instruction might need to be done in a pre-existing course for pre-service teachers that would likely lack hypothesis testing.

Conclusions

Variability is a very important part of understanding and using statistics. Being able to understand variability so that students can reason in a statistical manner is a goal for all students in K-12, let alone the students taking introductory statistics courses at the college-level. This study looked at a very small and conceptual part of this idea, students' reasoning about variability when comparing distributions without focusing on calculations. This skill is one people often use when reading an article with graphs or seeing a graphical display in a presentation. This study showed that students were able to reason about variability, some more correctly than others. However, even students who were incorrect were still reasoning about variability, not just a measure of center.

This study also found that students' reasoning about variation when comparing distributions changed as they progressed through an introductory college-level statistics course in several different ways, including improvement, maintenance, decline, and inconsistency. Overall, quantitative data showed no statistically significant change over the course. In addition, it was seen that, while students were good at reasoning about variability when using bar graphs and dot plots, they struggled with reasoning when using histograms. This adds to the existing literature by focusing on the progress made by college students in reasoning about variability, which has not been studied in depth. In addition, it gives examples of students' reasoning at higher levels that was lacking in previous studies.

References

Abdelbaki, A. (2012). *The teaching and learning of abstract and concrete mathematical structures at the college undergraduate level*. (Doctoral dissertation, University of Virginia). Retrieved from http://gradworks.umi.com/35/30/3530507.html

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

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology*, *25*(4), 527−547.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*(2), 42−63. Retrieved from http://www.stat.auckland.ac.nz/serj

Ben-Zvi, D., & Sharett-Amir, Y. (2005, July). How do primary school students begin to reason about distributions. In *Reasoning about Distribution: A collection of studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-4), Auckland*, 2−7.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. *Intelligence: Reconceptualization and Measurement*, 57−76.

Blair, R. M., Kirkman, E. E., Maxwell, J. W. (2013). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2010 CBMS survey*. United States: American Mathematical Society.

Canada, D. (2006). Elementary pre-service teachers' conceptions of variation in a probability context. *Statistics Education Research Journal, 5*(1), 36−63. Retrieved from http://www.stat.auckland.ac.nz/serj

Chan, S. W., & Ismail, Z. (2012). Assessing misconceptions in reasoning about variability among high school students. *Procedia-Social and Behavioral Sciences, 93*, 1478−1483.

Chance, B. L., & Rossman, A. J. (2001). Sequencing topics in introductory statistics: A debate on what to teach when. *The American Statistician, 55*(2), 140−144.

College Board (2011). AP Examination volume changes. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/AP-Exam-Volume-Change-2011.pdf

Common Core State Standards Initiative. (2010). Common core state standards for mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

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.

Cooper, L.L., & Shore, F. S. (2010). The effects of data and graph type on concepts and visualizations of variability. *Journal of Statistics Education, 18*(2), 1−16.

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches, 2nd edition*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods Approaches, 4th edition.* Los Angeles, CA: Sage Publications.

Dabos, M. G. G. (2011). *Two-year college mathematics instructors' conceptions of variation* (Doctoral dissertation, University of California).

delMas, R. C. (2005). A comparison of mathematical and statistical reasoning. In D Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79−95). The Netherlands: Kluwer Academic Publishers.

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. Retrieved from http://www.stat.auckland.ac.nz/serj

DeVeaux, R. D., Velleman, P. F., & Bock, D. E. (2014). *Intro Stats, 4th ed.* London: Pearson/Addison-Wesley.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-k-12 curriculum framework.* Alexandría, VA: American Statistical Association. Retrieved from http://www.amstat.org/Education/gaise/GAISEPreK-12_Full.pdf

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin, 105*(3), 331−351.

Garfield, J., & Alhgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, *19*(1), 44−63.

Garfield, J., Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., ... Witmer, J. (2005). *Guidelines for assessment and instruction in statistics education (GAISE): College report.* Alexandria, Virginia: The American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/GAISECollege.htm

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92−99. Retrieved from http://www.stat.auckland.ac.nz/serj

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*(3), 372−396.

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

Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal, 3*(2), 7−16. Retrieved from http://www.stat.auckland.ac.nz/serj

Hjalmarson, M.A. (2007). Engineering students designing a statistical procedure for quantifying variability. *Journal of Mathematical Behavior, 26*(2), 178−188.

Hjalmarson, M. A., Moore, T. J., & delMas, R. (2011). Statistical Analysis when the data is an image: Eliciting student thinking about sampling and variability. *Statistics Education Research Journal, 10*(1), 15−34. Retrieved from http://www.stat.auckland.ac.nz/serj

Inzuna, S. (2006). Some conceptions and difficulties of university students about variability. In S. Alatorre, J. L. Cortina, M. Saiz, & A. Mendez (Eds.), *Proceedings of the 28th annual meeting of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 244–250). Merida, Mexico: Universidad Pedagogica Nacional.

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.

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

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013). On the misinterpretation of histograms and box plots. *Educational Psychology, 33*(2), 155−174.

Makar, K., & Confrey, J. (2005). "Variation-talks": Articulating meaning in statistics. *Statistics Education Research Journal, 4*(1), 27−54.

Meletiou, M., & Lee, C. (2002). Students' understanding of histogram: A stumbling stone to the development of intuitions about variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society (ICOTS6), Cape Town, South Africa.* Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/1/10_19_me.pdf

Mevarech, Z. R. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*(4), 415−429.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook, Second Edition.* Thousand Oaks, CA: Sage Publications.

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

National Council of Teachers of Mathematics. Commission on Standards for School Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: The Council.

Noll, J. (2011). Graduate teaching assistants’ statistical content knowledge of sampling. *Statistics Education Research Journal*, *10*(2), 48−74.

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.

Noll, J., Shaughnessy, M., & Ciancetta, M. (2010, July). Students’ statistical reasoning about distribution across grade levels: A look from middle school through graduate school. In *Proceedings of the Eighth International Conference on the Teaching of Statistics (ICOTS8), Ljubljana, Slovenia.* Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_8B4_NOLL.pdf

Onwuegbuzie, A. J., & Leech, N. L. (2007). A call for qualitative power analyses. *Quality & Quantity, 41*(1), 105−121.

Peters, S. A. (2011). Robust understanding of statistical variation. *Statistics Education Research Journal, 10*(1), 52−88. Retrieved from http://www.stat.auckland.ac.nz/serj

Peters, S. A. (2013). Developing understanding of statistical variation: Secondary statistics teachers' perceptions and recollections of learning factors. *Journal of Mathematics Teacher Education, 17*(6), 1−44. doi: 10.1007/s10857-013-9242-7

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning*, *5*(2&3), 131–156.

Pfannkuch, M. (2005). Thinking tools and variation. *Statistics Education Research Journal*, *4*(1), 83-91. Retrieved from http://www.stat.auckland.ac.nz/serj

Reading, C. (2004). Student description of variation while working with weather data. *Statistics Education Research Journal, 3*(2), 84−105. Retrieved from http://www.stat.auckland.ac.nz/serj

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 201−226). Netherlands: Kluwer Academic Publishers.

Reid, J., & Reading, C. (2008). Measuring the development of students' consideration of variation. *Statistics Education Research Journal*, *7*(1), 40−59. Retrieved from http://www.stat.auckland.ac.nz/serj

Rossman, A. J., & Chance, B. L., (2008). *Workshop statistics: Discovery with data, 3rd edition*. New York, NY: Key Curriculum Press.

Salkind, N. J. (2009). *Exploring research, 7th edition.* Upper Saddle River, NJ: Pearson Prentice Hall.

Sanchez, E., Borim da Silva, C., & Coutinho, C. (2011). Teachers' understanding of variation. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics- Challenges for Teaching and Teacher Education: A joint ICMI/IASE Study (pp. 211-221).* The Netherlands: Springer.

Scheaffer, R. L., & Stasny, E. A. (2004). The state of undergraduate education in statistics: A report from the CBMS 2000. *The American Statistician, 58*(4), 265−271.

Sharma, S. (2007). Exploring pre-service teachers' understanding of statistical variation: Implications for teaching and research. *Australian Senior Mathematics Journal, 21*(2), 31−43.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465−494). New York: MacMillan.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second Handbook of Research on the Teaching and Learning of Mathematics: A Project of the National Council of Teachers of Mathematics* (pp. 957−1009). United States of America: Information Age Publishing, Inc.

Shaughnessy, J. M., Canada, D., & Ciancetta, M. (2003, July). Middle school students' thinking about variability in repeated trials: A cross-task comparison. In N. Pateman, B. Dougherty, & J. Zilliox (Eds.), *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education held jointly* *with the 25th Conference of PME-NA, Vol. 4*, (pp. 159–165). Retrieved from http://files.eric.ed.gov/fulltext/ED500860.pdf#page=171

Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a Statistically Literate Society, Cape Town, South Africa*. Retrieved from http://iase-web.org/documents/papers/icots6/6a6_shau.pdf

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (2004, July). Types of student reasoning on sampling tasks. In *Proceedings of the 28th International Group for the Psychology of Mathematics Education, Vol. 4,* (pp. 177–184).

Shaughnessy, J.M., & Pfannukuch, M. (2002). How faithful is old faithful? Statistical thinking: A story of variation and prediction. *Mathematics Teacher*, *95*(4), 252−259.

Shaughnessy, J. M., & Zawojewski, J. S. (1999). Secondary students' performance on data and chance in the 1996 NAEP. *Mathematics Teacher, 92*(8), 713−18.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114−145.

Slauson, L. V. (2008). *Students' conceptual understanding of variability.* (Doctoral dissertation, The Ohio State University). Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/dissertations/08.Slauson.Dissertation.pdf

Turegun, M. (2011). *A model for developing and assessing community college students' conceptions of the range, interquartile range, and standard deviation.*

(Doctoral Dissertation, University of Oklahoma). Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/dissertations/11.Turegun.Dissertation.pdf

Turegun, M., & Reeder, S. (2011). Community college students' conceptual understanding of statistical measures of spread. *Community College Journal of Research and Practice*, *35*(5), 410−426.

Torok, R. (2000). Putting the variation into chance and data. *Australian Mathematics Teacher, 56*(2), 25−31.

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

Tversky, B. (1997). Cognitive principles of graphic displays. In *Proceedings of the AAAI 1997 Fall Symposium on Reasoning with Diagrammatic Representations,* (pp. 116–124). Retrieved from http://www.aaai.org/Papers/Symposia/Fall/1997/FS-97-03/FS97-03-015.pdf

Watson, J. M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics*, *47*(3), 337–372.

Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics, 51*(3), 225–256.

Watson, J. M. (2009). The influence of variation and expectation on the developing awareness of distribution. *Statistics Educations Research Journal, 8*(1), 32−61. Retrieved from http://www.stat.auckland.ac.nz/serj

Watson, J. M., Callingham, R. A., & Kelly, B. A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning, 9*(2), 83−130.

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

Watson, J. M., & Kelly, B. A. (2002). Can grade 3 students learn about variation. In *Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa*, (pp. 7−12). Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/1/2a1_wats.pdf

Watson, J. M., & Kelly, B. A. (2004). Statistical variation in a chance setting: A two year study. *Educational Studies in Mathematics, 57*(1), 121−144.

Watson, J. M., & Kelly, B. A. (2005). The winds are variable: Student intuitions about variation. *School Science and Mathematics*, *105*(5), 252−269.

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

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

Watson, J., & Pereira-Mendoza, L. (1996). Reading and predicting from bar graphs. *Australian Journal of Language and Literacy, 19*(3), 244−258.

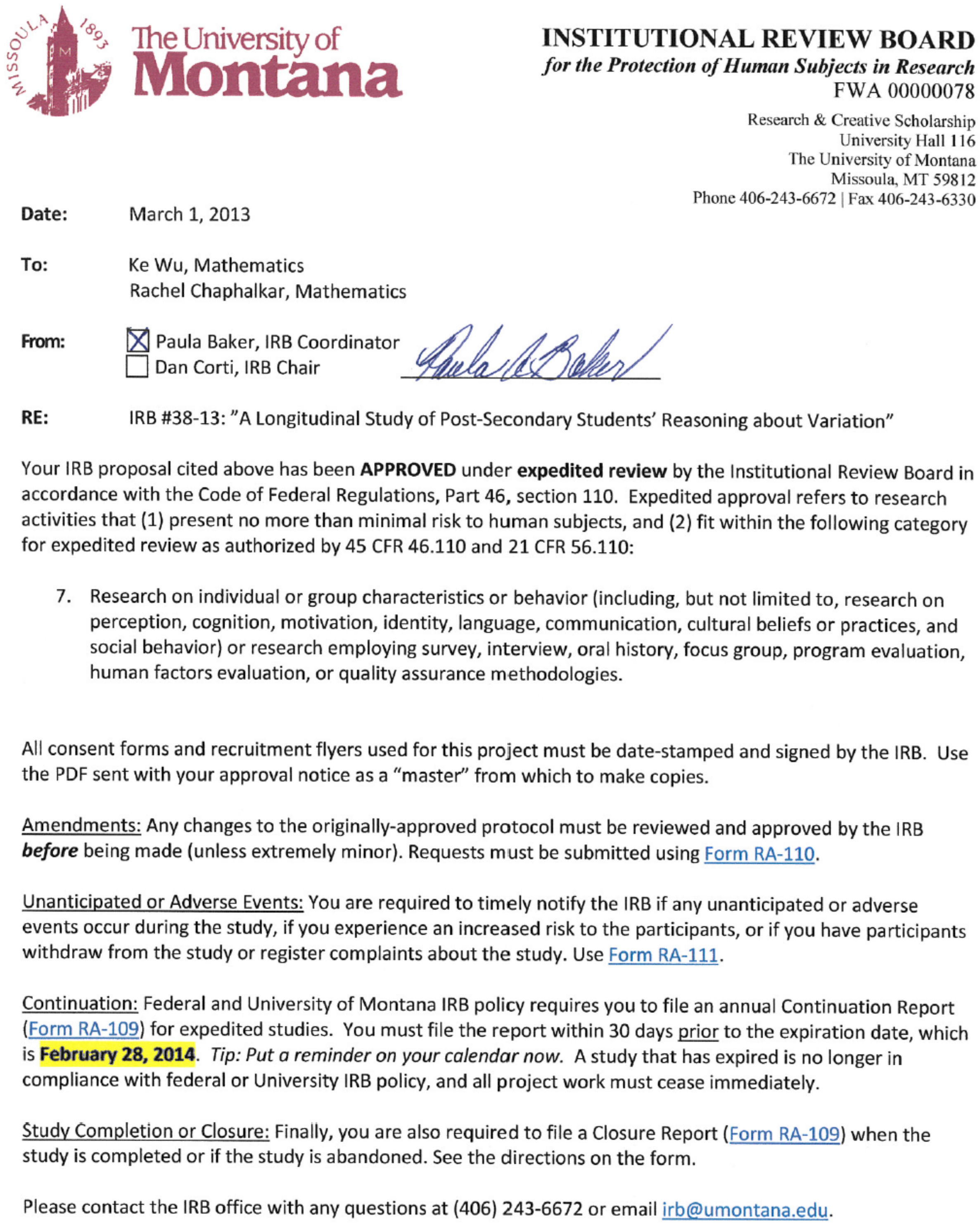
Watson, J. M., & Shaughnessy, J. M. (2004). Proportional reasoning: Lessons from research in data and chance. *Mathematics Teaching in the Middle School, 10*(2), 104−109.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal, 5*(2), 10−26.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223−265.

Zieffler, A. S., & Garfield, J. B. (2009). Modeling the growth of students' covariational reasoning during an introductory statistics course. *Statistics Education Research Journal, 8*(1), 7−31. Retrieved from http://www.stat.auckland.ac.nz/serj

Appendix A. Original IRB Approval



The University of Montana

MISSOULA 1893

INSTITUTIONAL REVIEW BOARD
for the Protection of Human Subjects in Research
FWA 00000078

Research & Creative Scholarship
University Hall 116
The University of Montana
Missoula, MT 59812
Phone 406-243-6672 | Fax 406-243-6330

Date: March 1, 2013

To: Ke Wu, Mathematics
Rachel Chaphalkar, Mathematics

From: ☒ Paula Baker, IRB Coordinator
□ Dan Corti, IRB Chair
Paula A Baker

RE: IRB #38-13: "A Longitudinal Study of Post-Secondary Students' Reasoning about Variation"

Your IRB proposal cited above has been **APPROVED** under **expedited review** by the Institutional Review Board in accordance with the Code of Federal Regulations, Part 46, section 110. Expedited approval refers to research activities that (1) present no more than minimal risk to human subjects, and (2) fit within the following category for expedited review as authorized by 45 CFR 46.110 and 21 CFR 56.110:

7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

All consent forms and recruitment flyers used for this project must be date-stamped and signed by the IRB. Use the PDF sent with your approval notice as a "master" from which to make copies.

Amendments: Any changes to the originally-approved protocol must be reviewed and approved by the IRB ***before*** being made (unless extremely minor). Requests must be submitted using Form RA-110.

Unanticipated or Adverse Events: You are required to timely notify the IRB if any unanticipated or adverse events occur during the study, if you experience an increased risk to the participants, or if you have participants withdraw from the study or register complaints about the study. Use Form RA-111.

Continuation: Federal and University of Montana IRB policy requires you to file an annual Continuation Report (Form RA-109) for expedited studies. You must file the report within 30 days prior to the expiration date, which is **February 28, 2014**. *Tip: Put a reminder on your calendar now.* A study that has expired is no longer in compliance with federal or University IRB policy, and all project work must cease immediately.

Study Completion or Closure: Finally, you are also required to file a Closure Report (Form RA-109) when the study is completed or if the study is abandoned. See the directions on the form.

Please contact the IRB office with any questions at (406) 243-6672 or email irb@umontana.edu.

Form RA-108
(Rev. 11/12)

THE UNIVERSITY OF MONTANA-MISSOULA
Institutional Review Board (IRB)
for the Protection of Human Subjects in Research
CHECKLIST / APPLICATION

IRB Protocol No.: 38-13

At The University of Montana (UM), the Institutional Review Board (IRB) is the institutional review body responsible for oversight of all research activities involving human subjects outlined in the U.S. Department of Health and Human Services' Office of Human Research Protection and the National Institutes of Health, Inclusion of Children Policy Implementation.

Instructions: A separate application form must be submitted for each project. IRB proposals are approved for no longer than one year and must be continued annually (unless Exempt). Faculty and students may email the completed form as a Word document to *IRB@umontana.edu*. or submit a hardcopy to the Office of the Vice President for Research & Development, University Hall 116. Student applications must be accompanied by email authorization by the supervising faculty member or a signed hard copy. *All fields must be completed. If an item does not apply to this project, write in: n/a.*

1. Administrative Information

Project Title: A Longitudinal Study of Post-Secondary Students' Reasoning about Variation	
Principal Investigator: Dr. Ke Wu	UM Position: Faculty (Associate Professor)
Department: Mathematics	Office location: Math 201
Work Phone: 243-4818	Cell Phone: (406)-274-2873

2. Human Subjects Protection Training *(All researchers, including faculty supervisors for student projects, must have completed a self-study course on protection of human research subjects* ***within the last three years*** *(http://www.umt.edu/research/complianceinfo/IRB/) and be able to supply the "Certificate(s) of Completion" upon request. If you need to add rows for more people, contact the IRB office for assistance.*

All Research Team Members (list yourself first)	**PI**	**CO-PI**	**Faculty Supervisor**	**Research Assistant**	**DATE COMPLETED Human Subjects Protection Course**
Name: Ke Wu Email: ke.wu@mso.umt.edu	☒	□	□	□	1/25/2013
Name: Rachel Chaphalkar (formerly Robertson) Email: rachel.chaphalkar@umontana.edu	□	☒	□	☒	1/27/2011
Name: Email:	□	□	□	□	
Name: Email:	□	□	□	□	

3. Project Funding *(If federally funded, you must submit a copy of the abstract.)*

Is grant application currently under review at a grant funding agency? □Yes *(If yes, cite sponsor on ICF if applicable)* ☒No

Has grant proposal received approval and funding? □Yes *(If yes, cite sponsor on ICF if applicable)* ☒No

Agency	Grant No.	Start Date	End Date	PI on grant

For UM-IRB Use Only

IRB Determination:

*** Note to PI:** Study is approved for one year only. Use any attached IRB-approved forms (signed/dated) as "masters" when preparing copies. If continuing beyond the expiration date, a continuation report must be submitted. Notify the IRB if any significant changes or unanticipated events occur. When the study is completed, a closure report must be submitted. Failure to follow these directions constitutes non-compliance with UM policy and will have consequences.

_____ Not Human Subjects Research
_____ Approved Exempt from Review, Exemption # ____ *(see memo)*
__X__ Approved by Expedited Review, Category # 7 *(see* ***Note to PI***)*
_____ Full IRB Determination
_____ Approved *(see* ***Note to PI***)*
_____ Conditional Approval *(see memo)* - IRB Chair Signature/Date: ____________
_____ Conditions Met *(see* ***Note to PI***)*
_____ Resubmit Proposal *(see memo)*
_____ Disapproved *(see memo)*

Risk Level: Minimal

Final Approval by IRB Chair/Coordinator: Paula A Baker Date: 3-1-13 Expires: 2-28-14

PARTICIPANT INFORMATION AND INFORMED CONSENT

Title: A Longitudinal Study of Post-Secondary Students' Reasoning about Variation

Project Director(s):

Ke Wu
Principal Investigator
Department of Mathematical Sciences
University of Montana
Missoula, MT 59812
ke.wu @umontana.edu
406-243-4818

Rachel Chaphalkar
Graduate Research Assistant
Department of Mathematical Sciences
University of Montana
Missoula, MT 59812
rachel.chaphalkar @umontana.edu
406-243-4486

Special instructions:

This consent form may contain words that are new to you. If you read any words that are not clear to you, please ask the person who gave you this form to explain them to you.

Purpose:

You are being asked to take part in a research study on student understanding of variation. The purpose of this research study is to explore how students' understanding of variation changes during an introductory statistics course.

Procedures:

If you agree to participate in this research study, you will be interviewed four times for approximately 20 minutes each time. Some interviews may be based on previously completed coursework. The interviews will take place either in the PI's office (Corbin 364) or in the mathematics conference room.

Risks/Discomforts:

There is no anticipated discomfort for those contributing to this study, so risk to participants is minimal. This study has no effect on your course grade and you have the right to withdraw from the study at any time, should you choose to participate. Your instructor will NOT have access to the interview audio recordings or transcripts.

Benefits:

You may better understand your learning of statistics by talking to the interviewer and you will help us to better understand student learning of statistics as a result of your participation.

Confidentiality:

Your identity will be kept confidential and only the researcher and faculty supervisor will have access to the files. Your signed consent form will be stored in a cabinet separate from the data. Once data are collected, your identity will be removed and replaced by a

The University of Montana IRB
Expiration Date 2-28-14
Date Approved 3-1-13
Chair/Admin Paula A Baker

code known only to the researcher. All data will be stored in a locked file cabinet and your signed consent form will be kept separate from the data.

If the results of this study are written in a scientific journal or presented at a scientific meeting, your name will not be used.
Your initials _________ indicate your permission to be identified by name in any publications or presentations.
If you do not want to be acknowledged by name in any publications or presentations, please initial here _________.

Interviews will be audio-recorded and the audio files will be transcribed without any information that could identify you. The tape will then be erased.

Voluntary Participation/Withdrawal:

Your decision to take part in this research study is entirely voluntary. You may refuse to take part in or you may withdraw from the study at any time without penalty or loss of benefits to which you are normally entitled.

Questions:

If you have any questions regarding your rights as a research subject, you may contact the Chair of the IRB through The University of Montana Research Office at 243-6670.

Statement of Consent:

I have read the above description of this research study. I have been informed of the risks and benefits involved, and all my questions have been answered to my satisfaction. Furthermore, I have been assured that any future questions I may have will also be answered by a member of the research team. I voluntarily agree to take part in this study. I understand I will receive a copy of this consent form.

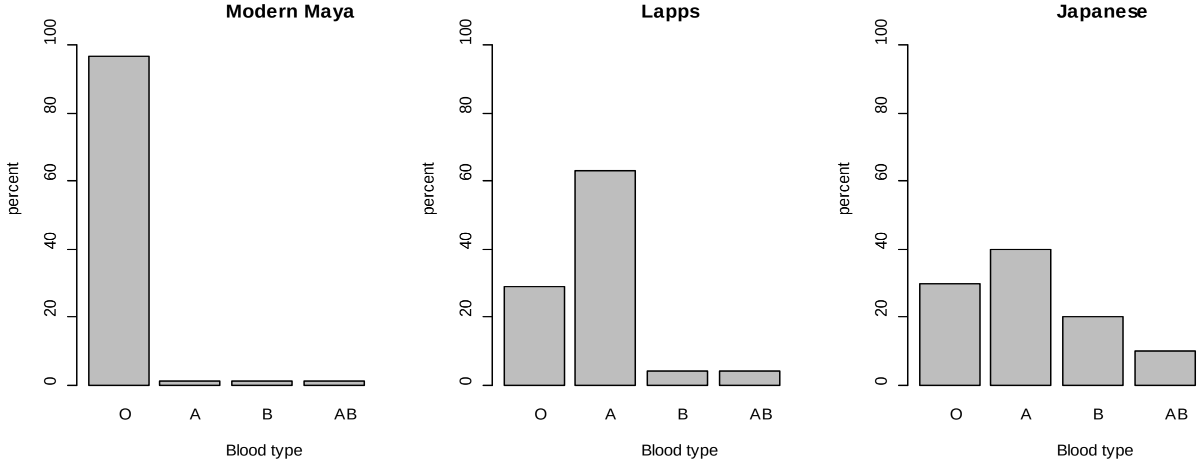
____
Printed (Typed) Name of Subject

______________________________ ______________________
Subject's Signature Date

The University of Montana IRB
Expiration Date 2-28-14
Date Approved 3-1-13
Chair/Admin Paula A. Baker

Appendix B. Online Survey 1

Question 1. Consider the distributions of the blood types of three different ethnic groups.



Without making any calculations, which group has the most variability in their blood types?

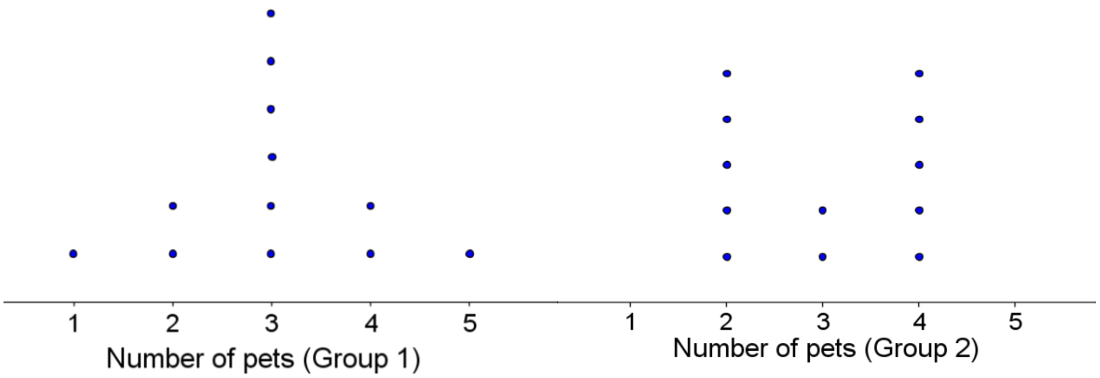
- □ Modern Maya
- □ Lapps
- □ Japanese
- □ No difference
- □ I'm not sure

Which group has the least variability in their blood types?

- □ Modern Maya
- □ Lapps
- □ Japanese
- □ No difference
- □ I'm not sure

Briefly explain your answer.

Question 2. Two groups of people were asked how many pets they own. Their responses can be seen in the following dot plots.

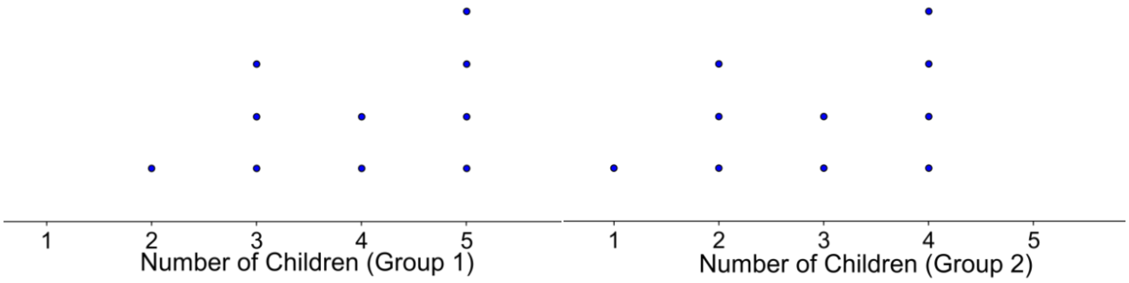


Which group has more variability in the number of pets they own?

- □ Group 1
- □ Group 2
- □ No difference
- □ I'm not sure

Briefly explain your answer.

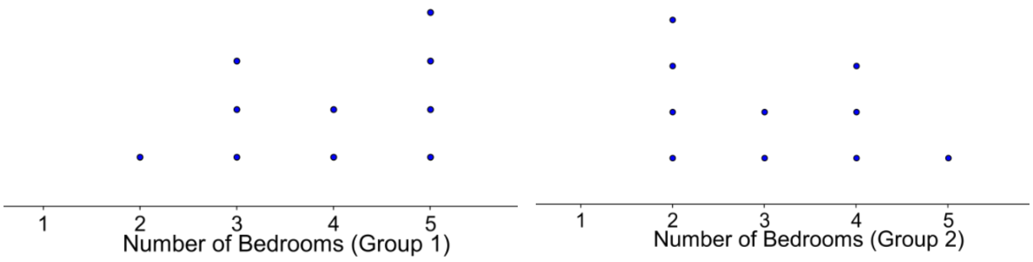
Question 3. Two groups of people were asked how many children were in the family in which they were considered a child. Their responses can be seen in the following dot plots.



Which group has more variability in the number of children in a family?

- □ Group 1
- □ Group 2
- □ No difference
- □ I'm not sure

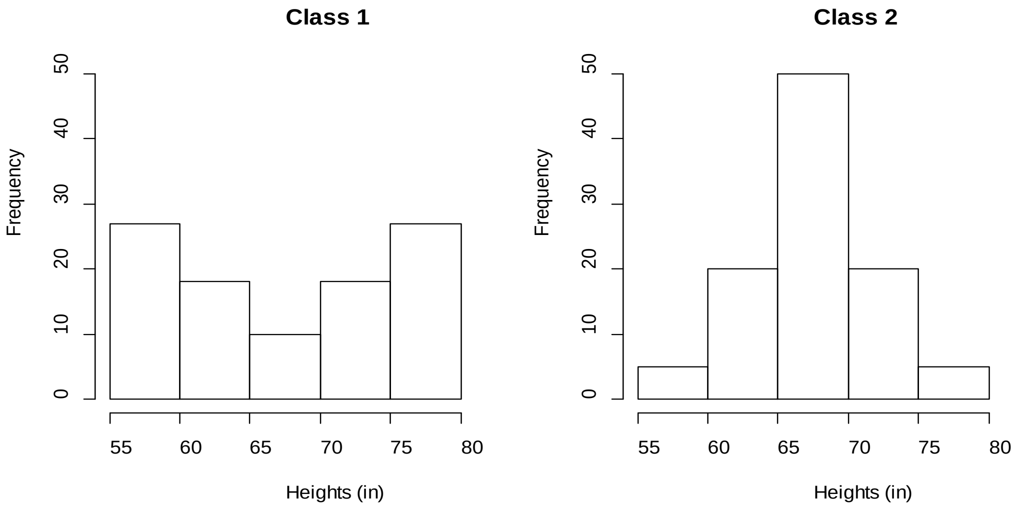
Question 4. Two groups of people were asked how many bedrooms are in the house that they currently live in. Their responses can be seen in the following dot plots.



Which group has more variability in the number of bedrooms in the house?

- □ Group 1
- □ Group 2
- □ No difference
- □ I'm not sure

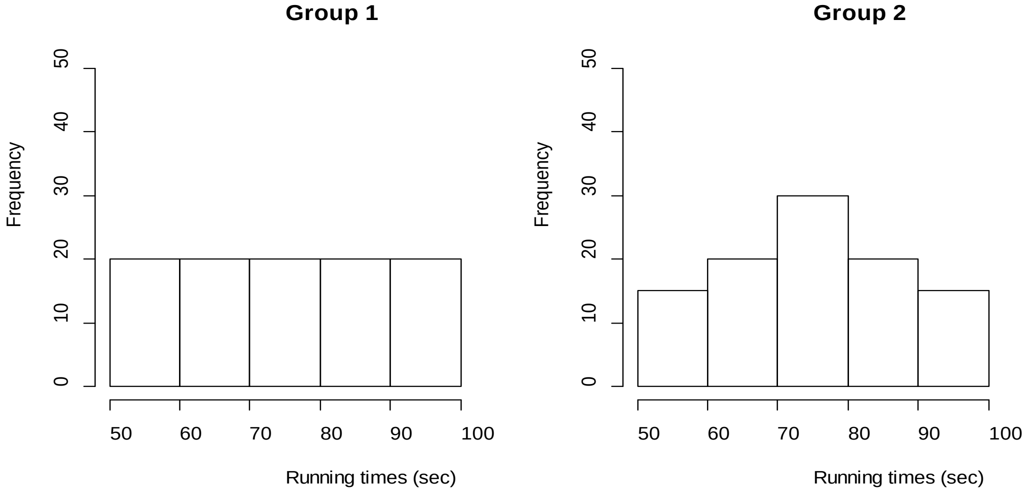
Question 5. Consider the distributions of heights (in inches) for two different classes.



Without making any calculations, which class has more variability in their heights?

- □ Class 1
- □ Class 2
- □ No difference
- □ I'm not sure

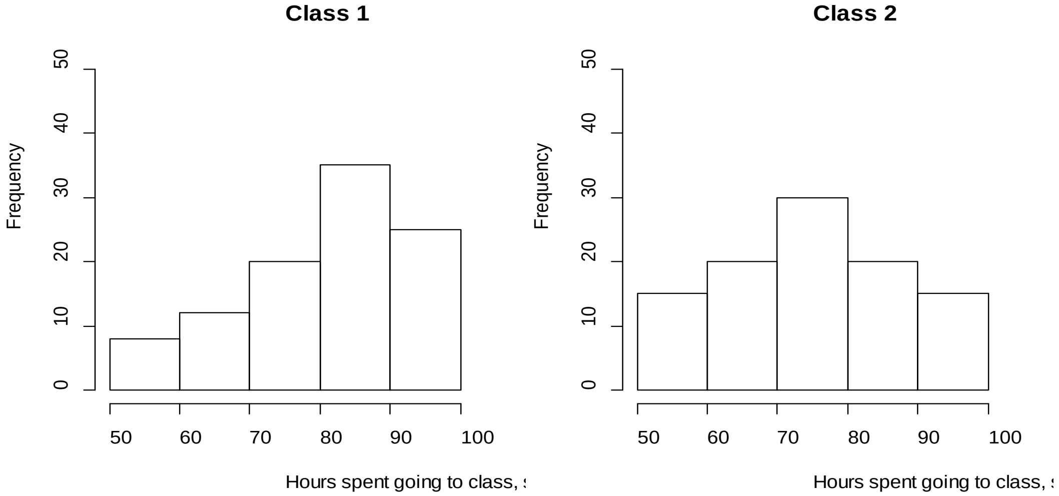
Question 6. Consider the distributions of time (in seconds) it took two different groups of runners to run 400 meters.



Without making any calculations, which group has more variability in their time to run 400 meters?

□ Group 1
□ Group 2
□ No difference
□ I'm not sure

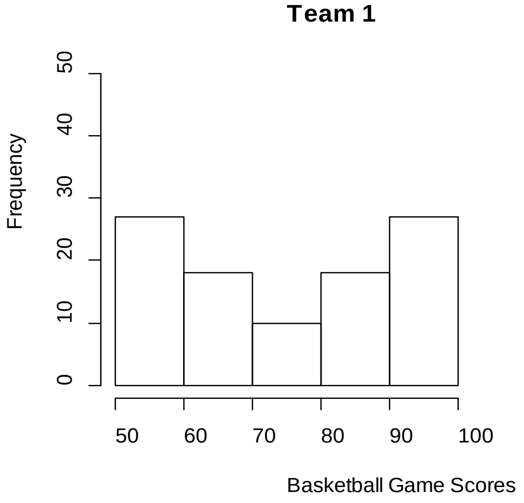
Question 7. Consider the distributions of time (in hours) it took two different classes spent going to class, studying, and working in one week.

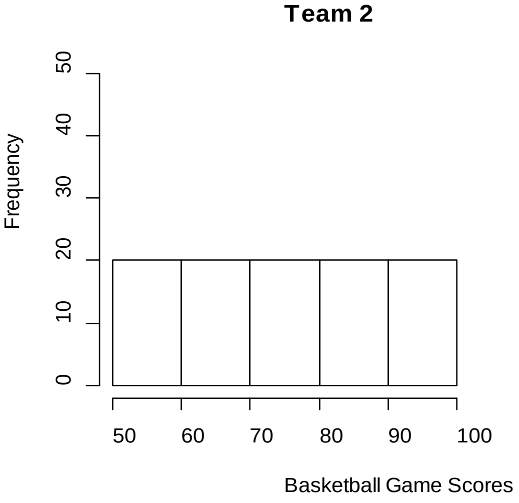


Without making any calculations, which class has more variability in their time spent going to class, studying, and working in one week?

□ Class 1
□ Class 2
□ No difference
□ I'm not sure

Question 8. Consider the distributions of points scored in a game by two particular basketball teams over 8 years.

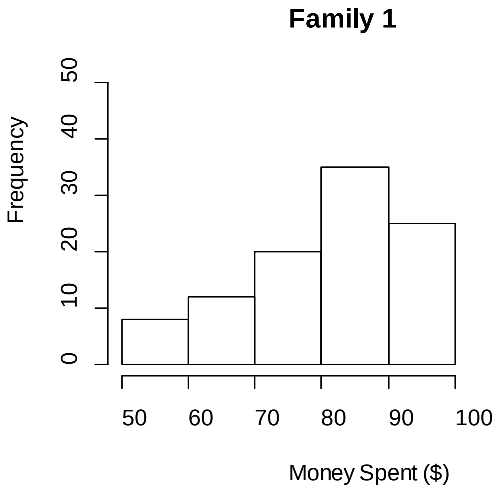


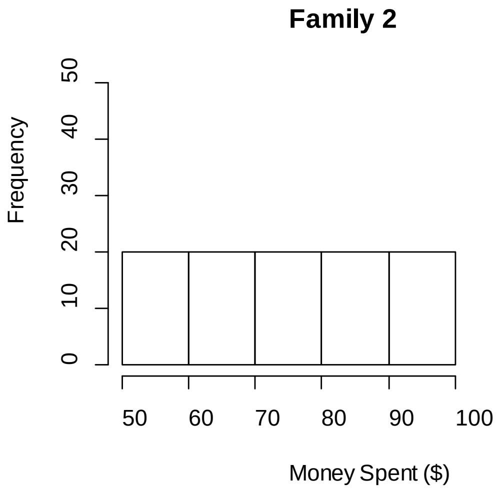


Without making any calculations, which team has more variability the points scored in a game?

- □ Team 1
- □ Team 2
- □ No difference
- □ I'm not sure

Question 9. Consider the distributions of the amount of money spent on groceries each week for two particular families over nearly two years.

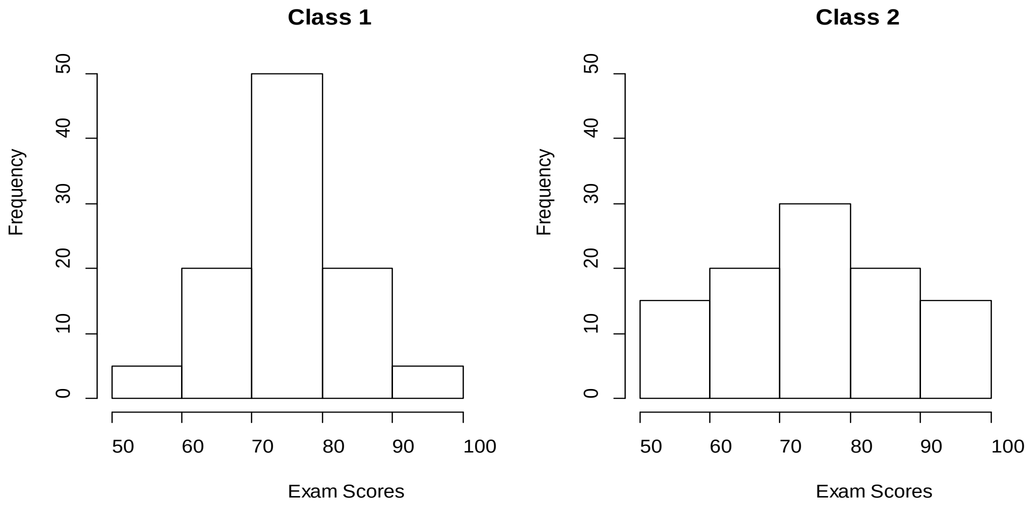




Without making any calculations, which family has more variability in the amount of money spent on groceries each week?

- □ Family 1
- □ Family 2
- □ No difference
- □ I'm not sure

Question 10. Consider the distributions of exam scores for two different classes.



Without making any calculations, which class has more variability in their exam scores?

- □ Class 1
- □ Class 2
- □ No difference
- □ I’m not sure

Briefly explain your answer.

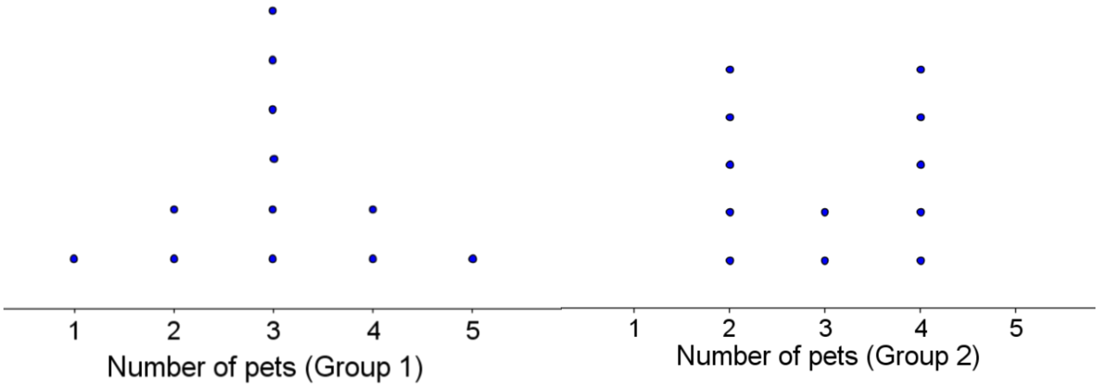
Appendix C. Interview Protocol 1

Have student sign informed consent before beginning. Answer any questions about informed consent.

Researcher: Thank you for agreeing to be interviewed. It will be very helpful if you can try to explain what you are thinking as much as possible. Sometimes we will look at some work you have already done for class such as Moodle questions or test questions. Please remember that what you say here will in no way affect your grade in the class.

Follow-up with student on his/her first answers to Online Survey 1, especially to clarify his/her explanations.

(Question 2.) Two groups of people were asked how many pets they own. Their responses can be seen in the following dot plots.

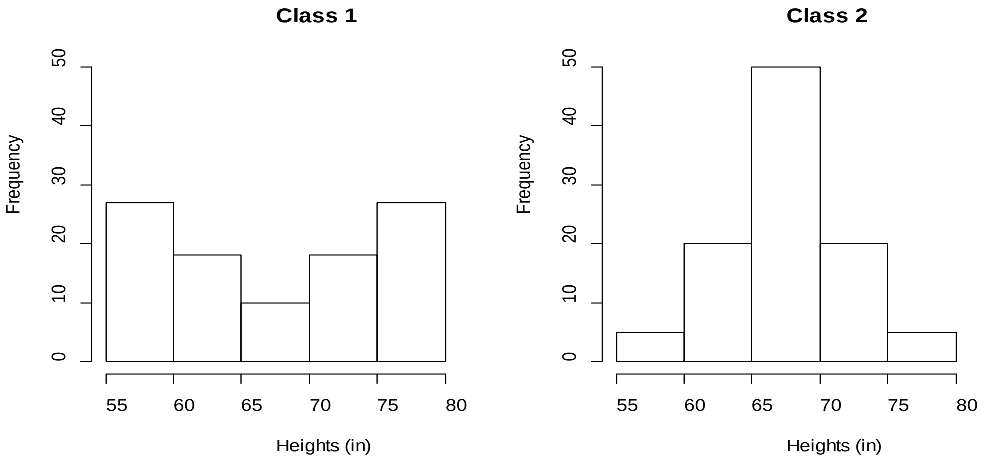


Which group has more variability in the number of pets they own?

Why?

If needed, ask students about different parts of the graph and why they are choosing their responses. Possibility get them to think about what sort of calculations could be made.

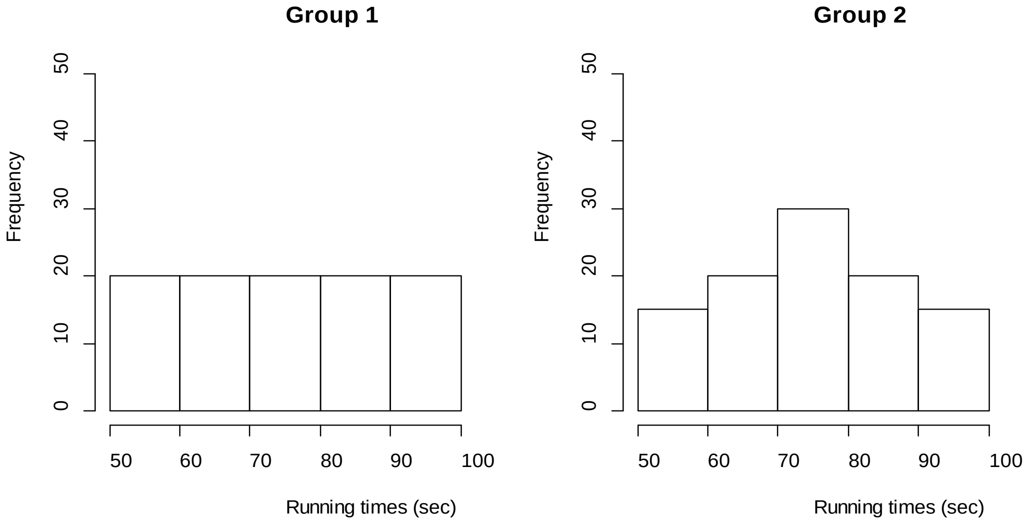
(Question 5.) Consider the distributions of heights (in inches) for two different classes.



Without making any calculations, which class has more variability in their heights?
Why?

If needed, ask students about different parts of the graph and why they are choosing their responses. Possibility get them to think about what sort of calculations could be made.

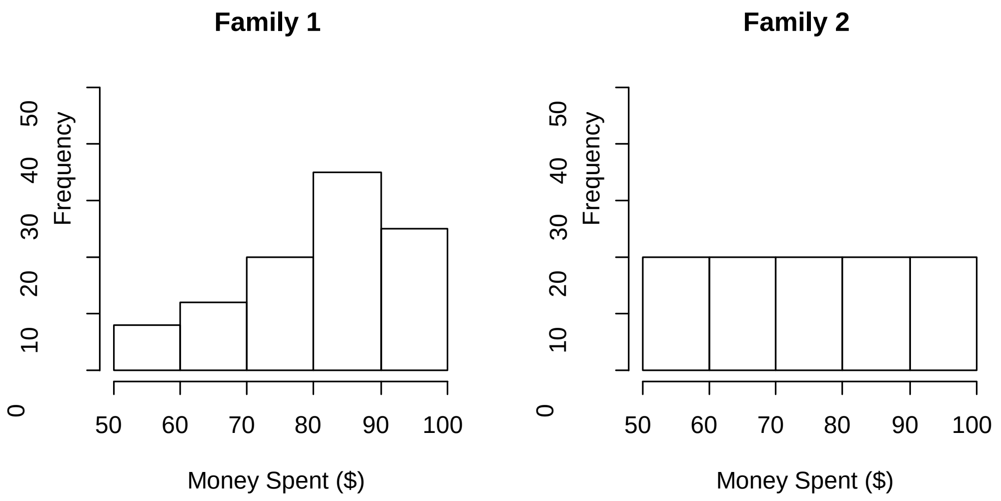
(Question 6.) Consider the distributions of time (in seconds) it took two different groups of runners to run 400 meters.



Without making any calculations, which group has more variability in their time to run 400 meters?
Why?

If needed, ask students about different parts of the graph and why they are choosing their responses. Possibility get them to think about what sort of calculations could be made.

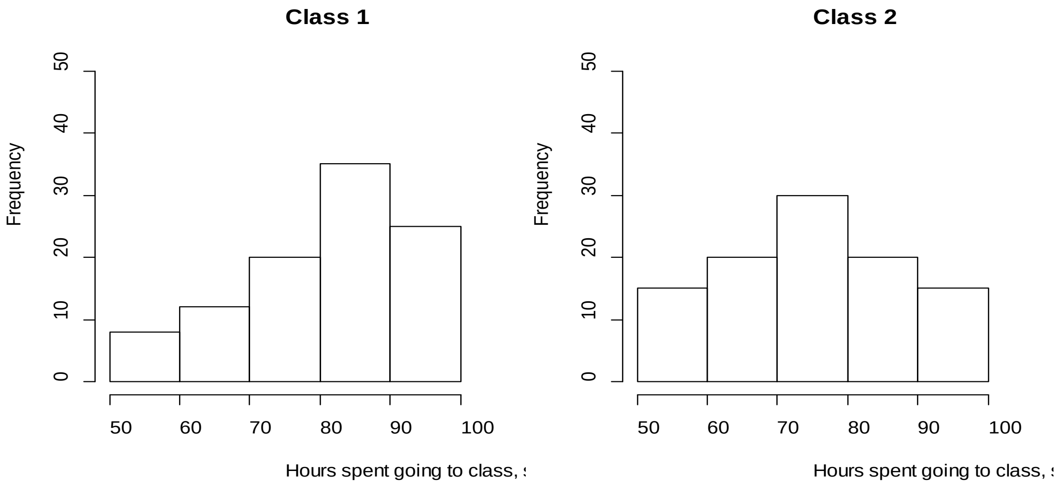
(Question 9.) Consider the distributions of the amount of money spent on groceries each week for two particular families over nearly two years.



Without making any calculations, which family has more variability in the amount of money spent on groceries each week?
Why?

If needed, ask students about different parts of the graph and why they are choosing their responses. Possibility get them to think about what sort of calculations could be made.

(Question 7.) Consider the distributions of time (in hours) it took two different classes spent going to class, studying, and working in one week.



Without making any calculations, which class has more variability in their time spent going to class, studying, and working in one week?
Why?

If needed, ask students about different parts of the graph and why they are choosing their responses. Possibility get them to think about what sort of calculations could be made.

In general, how do you decide histogram, dot plot, or bar graph has the most variability?
Prod student to explain as much as possible.

At the end of interview, thank student for their time and remind them they will be contacted for two more interviews.

Appendix D. Demographic Information

Please answer the following questions to the best of your knowledge. Please skip any questions you feel uncomfortable answering. Please circle your answers and choose all that apply.

1. What lab section are you in?

Lab 1 at 8am (9am lecture)
Lab 2 at 9am (9am lecture)
Lab 3 at 10am (9am lecture)
Lab 4 at 11am (9am lecture)
Lab 5 at 12pm (9am lecture)
Lab 6 at 1pm (9am lecture)
Lab 7 at 2pm (9am lecture)
Lab 8 at 3pm (9am lecture)
I don't remember my lab section (9am lecture)

Lab 11 at 8am (10am lecture)
Lab 12 at 9am (10am lecture)
Lab 13 at 10am (10am lecture)
Lab 14 at 11am (10am lecture)
Lab 15 at 12pm (10am lecture)
Lab 16 at 1pm (10am lecture)
Lab 17 at 2pm (10am lecture)
Lab 18 at 3pm (10am lecture)
I don't remember my lab section (10am lecture)

2. What is your current major? ____

3. What is your class standing?
Freshman Sophomore Junior Senior Post baccalaureate
Graduate student Other (please specify): ________________________

4. Previous college mathematics/statistics courses completed (either at the University of Montana or the equivalent at another institution):
Intermediate Algebra (M095)
Finite and Linear mathematics (M115)
Introduction to Statistics (STAT 216)
College Algebra, College Trigonometry, or Precalculus (M121/M122/M151)
Applied Calculus or Calculus (M162/M171)
None
Other (please specify): _______________________________________

5. Topics covered in your high school (grades 9-12) mathematics classes:
Algebra Geometry Trigonometry Precalculus Calculus Discrete Mathematics Statistics
Other (please specify):___________________________

6. Number of years of high school (grades 9-12) mathematics courses: ______

7. Sex:
Female
Male
Intersex

8. How do you currently pay for school?
I currently have federal student loans
I currently have private student loans
I have no loans

9. What is your age? _______

10. Your ethnicity (choose all that apply):
White
Native American or native Alaskan
Hispanic or Latino
Black
Asian
Native Hawaiian or Pacific Islander
Other (please specify): _________________

11. Current G.P.A. (please estimate if you are unsure):
3.5-4.00
3.0-3.49
2.5-2.99
2.0-2.49
1.5-1.99
1.0-1.49
0.5-0.99
0.0-0.49
This is my first semester; I do not have a G.P.A.

Appendix E. Selected Interview Transcripts

Brian Interview 1

Interviewer: So I think you’ve actually already answered this question, actually, you’ve answered all of these on moodle.

Brian: Yeah.

I: But, just to have some more idea, this is about the number of pets and two groups of people and which one would you say has more variability and why?

B: Uh, I would say that group 1 has more variability, just because it’s more spread out

I: Okay, and are you thinking about the 1 and the 5, is that?

B: Yeah, there is a lot in the middle but there’s still some here in the 2 and the 4 and then a little bit on the 1 and 5, versus this one, it’s just, there’s a lot in 2 and 4 and then just a little bit in 3 and that’s kind of it. If that makes sense.

I: Okay, no, that definitely makes sense. Okay, that’s fine, we can do a different one. Okay, so these ones are different about the heights of two different classes. These are histograms instead of dot plots, I guess, is there anything right away that you kind of… that changes how you’re thinking about it based on the type of display?

B: Yeah, definitely, I think now like this one, class 1 has more variability than here, because it looks like there is a lot more in the middle and less there, I don’t know, it just looks more spread out this time than before, this, maybe if it was over a little bit more, I don’t know, but yeah this one looks like there is more variability because you have a bigger range that is spread out, I don’t know, the columns all look pretty close in height and they’re all kind of spread out versus everything in the middle looks, if that’s like where the bulk of it is…

I: So more of the data?

B: More evenly distributed than this one.

I: Than class 2?

B: Yeah

I: And then I guess is, is there any difference in the type of information that you get, going from the dot plot to the histogram that makes it easier or harder to make a decision, because these are kind of similar, right, group 1 kind of goes with class 2 and group 2 kind of goes with class1, not exactly but…

B: Yeah pretty similar, yeah I guess just looking at it, maybe because this one has 5 columns and this only has 3 columns, maybe that is one of the distinguishers that I look at, I mean, obviously, group 1 still has 5 columns like class 2 but group 2 doesn't have the same amount of columns as class 1, so when I think about it, I'm like since class 1 has 5 columns, it looks like it might be more spread out.

I: Okay, good, okay, so you know, lots of very similar sorts of graphs, so how long it took people to run 400m, clearly made up (laughs)

B: Yeah, no, this would be interesting to see everyone has the same amount of running, although I feel like this one would have more variability, because it's completely distributed across, versus a lot more in middle versus the sides

I: For group 2. What do the running times look like for people in class 1?

B: They all look like they ran for, well, they all look like they ran for the same amount of time, yeah, because each person ran for… well, there were 20 people in each group of time frames that ran for the same amount, like 20 people ran for 50, and 20 people ran for 60 and 20 people ran for 70, which is kind of odd that is all evened out, yeah

I: Okay, so, and yeah, I know you said 50, so do you mean that for those 20 people they were all at exactly 50 seconds?

B: Um, between 50 and 60

I: Just somewhere in there

B: Yeah, I guess, yeah, as just 50, that would be kind of strange, but I remember that 50 to 60 doesn't just mean 50 and 60, like, every number from 50 to 59, which kind of means that it kind of is more distributed, a lot of people ran between 70 and 79 seconds versus these guys on the ends, there were less people that ran for those amounts.

I: Okay, so you're saying then that group 2 then has less variability because you have more people that ran in those middle times?

B: Yeah but maybe I am thinking about variability differently.

I: How are you thinking about variability?

B: Um, like distributed across, there's… it's more I don't know, distributed, it's more equally distributed, there's more variability because it's more evenly distributed, versus here in this middle column, it's…

I: There are more people.

B: There are more people in that part, so that means if there are more people in the middle, there are less people in the other two, so that there is less variability.

I: No, I think variability is a hard word to define; it's kind of a feeling in some ways. Okay, good. Okay, so uh, kind of similar but a little bit different looking distributions of these families, you still have the one where they are very evenly distributed, and then one that looks somewhat skewed.

B: I don’t know, I still think family 2, cause it’s…

I: Has more variability?

B: Yeah because there’s a lot more down here at the bottom versus here, there’s a lot at the top and a little bit between 50 and 60, so, versus this, there are a whole bunch of different ranges that are about even.

I: Sure, do you think it’s possible to like make up a data set, where maybe these pictures aren’t a very good representation of where the data actually are, still have the same frequencies in each, but do you think you could kind of cookbook it so that it?

B: yeah, I don’t know, if you made smaller bins, I think, you might actually be able to see that there’s actually 10 people that all have 80 to 82 versus between 85 and 89 there are only 8 people, so that if you look at it like this, you don’t catch that difference, you know it could be everybody could all be at 85, so there wouldn’t be that much variability, where here something like that but…

I: Good. Okay, one last one.

B: Hmm, well, I would probably say class 2, but depending on if there were smaller bins maybe, there would be, probably could be a bunch in class 1 that are a lot of different ranges here, between 80 to 89 versus here there could be a bunch of ones at 75, so looking at that, there might not be as much, but that could be the same case here in class 1, versus in class 2, but I still stick to that, this seems to be more evenly distributed, so, there’s more potential variability

I: Okay, so you feel like class 2 has more variability than class 1?

B: Yeah, so, without knowing what the actually data is.

I: It’s a hard judgment call, right, to decide on numbers. Okay, so when you’re answering these questions, did you thinking at all about measures of spread that we’ve discussed in class? I feel like on the first question, you were really thinking about range. Was there anything else that came in?

B: Yeah, um, I thought… I really, I’m really visual, so when I look at these things, I see, this one is really more evenly distributed, so I guess I would think about the range a lot, looking at the whole picture instead of narrowing down, like you said, well, what if all of these are the same, that would potentially mean that there wouldn’t be as much variability in one of the columns, so, but, coming from me who is an art person, I look at it and this looks all evenly distributed versus this, maybe if the bins were smaller, I would think more about the numbers and everything else.

I: Yeah in the dot plots, they were smaller, so, right, that kind of, you know exactly where each of those numbers are.

B: Yeah, so.

I: Good, that sounds good.

Brian Interview 2

Interviewer: Okay, so, have you gotten a chance to do the online questions?

Brian: Yeah, I already did it, the second survey?

I: Yeah

B: The same questions, sort of.

I: Sort of

B: Sort of

I: Sort of, but different context, huh?

B: Yeah, so

I: There's something to that, right, and yes, there's something to it, it does seem like some people are very motivated by different context. Do you feel like you're? With biology?

B: Yeah, I think so. Yeah, I like to think about, well, there might be some numbers presented but…

I: What does it mean?

B: What does the relativity of them to…? Because a high number might not necessarily be a good thing in one context but

I: It might not be a bad thing in another. Good point, well, you've looked at these, but then we'll do this. Different types of pets owned by people in different cities, Missoula, Bozeman, Billings, which one would you say has the most variability in the number of pets?

B: Definitely, I'd have to say Billings, it definitely wasn't Missoula, because there were only dogs, but you know, if you think about different kinds of dogs, then there might be some variability in dogs.

I: Yeah, sure. And the kinds… well… let's not make it too complicated

B: Yes, so but this one definitely, there's a lot more variation you have some people with dogs, some people with cats, some people with rodents, some people with reptiles, versus Missoula which is just people with dogs and very minimal, if any, of the other types of pets and then Bozeman is mostly dogs and cats.

I: Okay, so it was easy to kind of put Bozeman in the middle because you had most of two things? And Billings had a lot of several

B: Yeah, yeah, and then Missoula only had dogs

I: So that was the least, yeah, okay, sick of that question. Okay, so how many times people donated blood.

B: You know, I still kind of think that this one has more variation.

I: Group 1?
B: Group 1. Um, just because it's spread out more. It took me, I thought about it a couple times, because I was like, ah well, you know, maybe some sort of variable that I didn't think about, that I could probably say, oh yeah, that one has more variation, but this one definitely is spread out more, even though there's this extra amount in the middle,

I: They're kind of clumped in that peak.

B: Yeah, so, but, yeah.

I: So, on this previous one, you kind of like, to kind of compare these two, so in Missoula, you had everybody kind of in one bar.

B: Right.

I: And then lots of different ones, so to me, this one bar kind of looks like this one bar in group one for the blood types.

B: Yeah, definitely.

I: So this one kind of looks like to me the dog and cat one, but is that a fair comparison to make? Does it? I mean, you felt really confident about these two, what's different that makes these kind of?

B: Yeah, I have to think that it's kind of what we had talked about before, with the number of dots, or how spaced the dots might be, I don't know um, cause yeah, this one might be really high and be similar to this, this one's not as spread out, and I don't know, is each one of these dots one person?

I: Yeah, sure, each one of these dots is one person in the group, and it's representing how many times they've actually donated.

B: So I think this one might actually have more but there's one person on each end of this range, spectrum, sort of.

I: Okay, okay, no that's okay, that's what I want to know. Okay, so, difference in arm spans in two different classes.

B: This one has more variability. Um, which is interesting because it's kind of like that.

I: It is kind of like that.

B: Um, but um, it's…

I: So if you flip-flopped from the last question to this question, what makes you feel okay?

B: Maybe it's the number of columns that are…

I: Okay, that they each have 5 columns.

B: So this one has 5 columns, but this one has five columns also, but they're more spread out with the five columns compared to this one, it's like, okay, there's still five columns but this one is actually using 5 columns, and this one only uses 3. So, um, yeah, that's probably what made me this differently about it.

I: Okay, okay.

B: Yeah because like, if this one had maybe gone out a little bit more,

I: So if you had just had a few dots, like the first one, would it be harder to decide if the range here was from 1 to 4 and

B: I would take this one.

I: If it had anything in the column for 1?

B: Even if it had 2 or 1 or 1 dot over there or something.

I: And you wouldn't even feel like you needed both?

B: Just one or the other, yeah.

I: Okay, for group 2 and then it would have more. Okay, and then this one felt a little morbid… age of grandparents' death…

B: This one, I think, has more variability

I: Class 1.

B: Yeah, because like evenly distributed across on the whole thing versus there's a little bit more here, of course it could be a good bunch of variability between those numbers

I: Between 70 and 80?
B: Between 70 and 90.

I: So it could be half were 71 and half were 79.

B: Right, right versus you know, here it could be the same thing, 70 to 80 is a big age range for deciding, I'm sure if you had them by 5's or something, it wouldn't necessarily be okay well, maybe there's more variability in the other one, so I'd pick this one.

I: But you would go with class 1

B: Yeah,

I: Kind of assuming it is an even distribution for each bar?

B: Yeah, yeah.

I: Okay, last one

B: Alright, I think this one.

I: Group 2 on Halloween costumes.

B: Because a lot more people are in group 1 are hogging up space in the middle, what is it, 70 dollars like 70 to 80 dollars on costumes, versus like there's a more even distribution of how money is spent in group 2.

I: Okay, so you don't have as many people spending from 70 to 80.

B: Yeah, there would, I would think of more variability with group 2 than group …

I: So on one of the other questions, you mentioned, that it's hard to tell because of the range. Could you make up numbers for these so that you'd be convinced that group 1 would have more variability than group 2? Do you think it's even possible?

B: Um, I think maybe,

I: Like you could tell people, "you have to spend this much, you have to spend that much" if you could do that, do you think you could even be possible for them to flip flop?

B: Um, maybe… like, I guess, if you had to tell these people in the middle to spend less than 75 or spend between 70 and 75 or the others 70 and 74, and 75 to 80, or 79, just trying to think of how to…

I: To tell them what to do, sure

B: Then you might see a lot more variation in that one

I: Especially if you were able to make more bars

B: Yeah

I: Think that would probably be…

B: More bars would definitely help on this one. Um, this one does fine with the 5 of the bars, but even you could see how many people spent more or less in each of those groups, then, I guess it would be helpful, to make a better assessment.

I: Right, because you have to assume a lot.

B: Yeah

I: Sure, sure. Thanks.

Brian Interview 3

Interviewer: So, take a look at that one.

Brian: Alright. (sigh) I thought about this one when I was doing that survey and I was like ah well, it could be the second one but I don't know, I still went with, I still feel like group 1 has more variability, but I think it's because of the dots, looking at the number of dots, and even though they don't necessarily mean like that, you know, whether there's one person or not, number of pets, I think, oh well, there's still like 3 dots over here and three dots over here and there's what? 6 dots in the middle, versus like here there's, yeah there's two here but then there's five on each side, and maybe if there was one here and one here… you know…

I: That would make it so group 2 would have more variability?

B: Yeah, yeah. So.

I: Was it a tough call between the two of them?

B: Yeah, I thought about it a little longer than I had done before. Just like aw well, yeah but this one, and if there were more dots like here or just over the whole thing then I might have thought this one because I don't know like, when I look at, oh well there's only 1,2,3, how many dots there are,

I: So you're saying like there's three dots on either side that aren't in the middle?

B: Yeah, there's like 12, yeah 12 dots total, and I was like okay, there's 6 in the middle, but then there's 6 on either side of the middle, where this one, it's not as spread out as much, maybe if there was one of the dots on each side, then I would say this one.

I: Okay, okay, okay, did you think about, I know it says don't make a computation, but did you think about any computations as you were doing it?

B: Um, not really, I mean, other than just adding up all the dots.

I: So really just…

B: Seeing how many dots there were and if they were spread out.

I: Okay, okay, so and, and, um, so, you're going with group 1 as having more variability, but um, are you thinking about that just based on the range or is there the position of the dots, I guess? Does that make sense? Is it really based on range?

B: I think I'm looking more on the range of it and the actual like, positioning.

I: So if you had moved these two dots in group 1 from 2 over to 1 and these two dots from 4 over to 5 so you just had 3 and then 6 and then 3, would that change your answer at all?

B: Yeah, I think it does. I think it might have. But…

I: I guess would that make it an easier call or a harder call?

B: I think it would make it a harder call because you would have something similar to this except kind of spread out more

I: Okay, on that one.

B: Alright, I definitely said this one

I: Class 1?

B: Yeah, which is funny, compared to that one.

I: Because you kind of did opposite. What makes you go straight for class 1?

B: I think because it's more bins that, that there's information in, where this one is one in 5 and there's not really anything in and here the same amount, the range is like the same but more equally spread out in class 1 than in class 2,

I: Ok, ok, and that one.

B: I said this one, group 1, just because it's completely evenly distributed. I don't know there would be more variability because there are you know, a bunch here, maybe the same amount that are here and the same amount here and the same amount here versus it still kind of looks sort of like this just not as high up.

I: Okay, so I guess is there a word that comes to mind when you look at either of those group 2's or a name for the shape, I guess?

B: Oh yeah, the unimodal, definitely, but just, the middle, the median, is just like huge in this one but in group 2, it's, this one's kind of a tough one the more that I thought about it, because I would think that this one, group 1, would have more variability because it's just equally distributed, versus this one there's less equal distribution, so I guess I would think that the more equally distributed, the more variability. I don't know if that's true or not, but that's the way I was looking at it.

I: Okay, okay.

B: Um, I would say this one.

I: Class 2?

B: Class 2. For the same reason, um because this one is now kind of skewed,

I: Yeah.

B: To the left, skewed left, yeah, and this one is more equally distributed. (chuckles)

I: Okay, okay, okay. I agree with you. one last one.

B: And I would say group 2, cause…

I: For the same reason?

B: Yeah, sure. For the same reason, I had to think about what I was thinking about when I was taking the actual survey, um, and yeah, this one just most of it is in the middle versus like here there's actually some spread, some variability, you know?

I: Sure, even though they have the same range

B: Yeah

I: You're still feeling?

B: Yeah, yeah, (laughs) the first one is like, like exactly the same as this one, but then obviously that one's not and yeah (laughs) this one.

I: Group 2. So when you're thinking about variability, it sounds like you're thinking about how spread out a distribution is. Do you have any?

B: Yeah, about how evenly it's distributed.

I: And that helps you find what has most, okay sure.

B: Yeah, because there's a ton here, you know, not as many there, but then okay, these might be the same, but then they're way more here than in the middle of this one, so like if this was more spread out like this one

I: So you kinds of want to take the middle bar, and if you end up kind of distributing it among other groups?

B: …break it in half and (laughs) or even, I don't know, even if you put some over here, I feel like more needs to go there too.

I: Uh huh

B: On the right side of it.

I: Okay, sounds good.

Emily Interview 1

Interviewer: So, you did some moodle questions and I have some of them from that survey here, so I’ll ask you some again, but this is the one about the pets, there are two dot plots and which would you say has more variability in the number of pets that they own?

Emily: I would say group 1.

I: And why did that one jump out at you?

E: Mostly because it has a larger range that the numbers fall into, mostly in the 3 category and it expands from 1 to 5 instead of just 2 to 4, so I just saw that there was a bigger range.

I: And did you have any, I mean, that with confidence, right? So it seems that when you looked at it, even though there are more in the 2 and 4 column, it didn’t seem too, seem as variable as the other?

E: Yeah.

I: Okay, so, histograms, right, and I kind of realized that these are kind of similar questions, right?

E: Uh huh.

I: The shape of the distributions, I mean, this doesn’t have more dots on those sides, but on these histograms of heights of students, which one would you say has more variability?

E: Oh gosh, I’m trying to remember what I put in the survey. I think I said class 1 because it’s a little bit more uniform than class 2, and so with more in each separate category, there is a greater chance of, how do I explain this, I guess there is a greater probability that you don’t know which one it is, versus class 2, where most of the data is in the 65 to 70, so it’s a good probability that your data is in there, so it’s less variable. That’s how I thought about it.

I: Okay, and um, let’s say, if it wasn’t a bar between 65 and 70, but if the bar was like right here between 75 and 80, if you just switched the third bar and 5th bar in class 2, would you still feel like class 1 is more variable?

E: Um, yeah, I think so, just because you still have, this is like double plus some, so greater proportion, it’s not as variable, which outcome you’re going to get you’re still a little unsure, there’s not one that’s sticking out among the rest.

I: Okay, so all of the bars are kind of similar in height.

E: Uh huh.

I: So, here are some that are very similar in height, compared to some a little bit taller, so this is the running times, and…

E: (laughs) I think for this one, I also said group 1, just because they are uniform.

I: So that's more variable?

E: Right, I don't know, it kind of the same thing, but not as drastic as the last one. There is still a greater proportion in here, um, I don't know…

I: So um, from a dot plot to a histogram, is there any information that you get in one that you don't get in the other that makes the question easier or harder to answer, I guess?

E: Um, I think you can definitely do this with a histogram, but in a dot plot, you get the individual values, we have 1, 2, 3, 4, whereas this is a group of numbers, 50 to 60, I think you could make a larger histogram and go 50 to 100 and do dots for that, but that would get excessive

I: Yeah, it would be hard to look at.

E: That actually might maybe, you might not have this normal distribution or this uniform distribution.

I: It wouldn't look exactly uniform? Uh huh, I agree. And uh, do you think, if you were just making up data, do you think that you could make up data that could into this uniform-looking histogram but might actually be less variable than this group 2 or do you think regardless of the numbers that you pick that it, this one will always be more, will always have more variability? So out of these 20 running times, you could make up any numbers that fit into this 50 to 60 category and any that fit in 60 to 70, so like for each of these, and the same thing for this, and you'd still have as many in each, do you think you could pick numbers and really make it look like this group 2 actually has more variability, if you picked the right numbers?

E: I think you could because you could do a bunch that are 51 and just have like all the 20 in the 51 category, probably a lot of these are a good range like from 51 to 59 or something, um, and same thing over here, if you had a good range, and didn't have it weighted to one side or the other, then it would probably start looking a little more uniform and this one would probably get skewed more.

I: Okay, okay, so another uniform distribution, compared to a different looking one in families.

E: Variability still?

I: Yep, it's the same question, time and time again, but they are different, right?

E: Yeah.

I: I, do you think that this question is harder than the last one or?

E: Yeah.

I: I think so too.

E: Just because it’s skewed off to one side. Yeah, I’m not too sure about this one. My instinct is saying family 2 again.

I: Has more variability?

E: More variability because you aren’t sure what the outcome would be on a randomly chosen point, but with this one, if you take half of the data, it’s probably going to lie between 80 and 100, so it’s less variable, that’s my logic.

I: No, no, I mean, that’s what I want, is your logic. I know, my logic isn’t as interesting. So the last one is two different classes and their time spent studying and working, and that skewed one again.

E: Right, more variable, these ones are always so hard, the first one was easy to tell, at least for me, it’s…

I: It’s more dramatic.

E: Yeah, variability to me signifies like, like the range almost. But then when these have both 50 to 100, it’s a lot more difficult to say which one is more variable, I’d probably say class 1 because you have these groups that are smaller, and then you have these really big groups um, and these ones are all hovering close to the same around, that sounds so completely opposite from what I just said.

I: No, that’s alright, okay, so you’re even though class 2 is a little bit mound-shaped, you’re saying that the bars really aren’t that different in height, so you’re almost thinking that class 2 is kind of uniform-ish? Is that what?

E: Yeah.

I: So a little bit but not quite.

E: Which the other one had uniform and I said uniform has more variability,

I: No, this is good.

E: I don’t know why this is so hard to decide.

I: You feel that class 1 is probably more variable,

E: Yeah.

I: Just because, um, oh boy, okay, so you kind of said the variability is the range, but on all of these, you haven't been able to use range except for that first one, so what was kind of your second tier of, if you couldn't look at range, then you look at?

E: I looked at the frequency of all the options and if, these ones were fairly close, but if there was one bar that stood out way above the rest, then…

I: So that made it less variable, because you're more likely to pick something in that bar, if you were randomly…

E: Yeah.

I: So, okay, so now, you use range on the first one, on the dot plot, does range kind of just, is that, is that the most important thing to look at and after that look at the bar heights, do bar heights with range even matter?

E: Range is the first thing I think of when I think of variability, um, so, yeah when I looked at this one, I didn't really look at the heights of the dots first, I just looked at how far they range.

I: Range, uh huh.

E: But I think, it's hard to say, if you took these two away.

I: The 1 and 5 away?

E: Yeah, then they have the same range,

I: Then which one would you say is more variable?

E: Gosh, I don't know, probably 2, at that point, yeah, it's, yeah, because then I would look at the bar heights and you have more dots on either end rather than in the middle, I think that would be…

I: But with that 1 and 5 you'd definitely go with that as being more variable?

E: Yeah.

Emily Interview 2

Interviewer: …distributions of the types of pets people owned and which one would have more variability in the type of pet that they own.

Emily: Oh gosh, I haven't thought about this…

I: Don't worry, take your time, you're definitely allowed thinking time.

E: I guess I would choose Billings because it's like more evenly distributed. Um, I don't know, so like, it's not quite sure. In Missoula, it's pretty clear that almost anybody you're going to ask will have a dog, but if you ask someone in Billings, it's not going to be a clear answer, so I would say they have more variability.

I: Ok, so, and Bozeman, do you feel like?

E: Yeah, I was going back and forth between these two, but since both of these are like really, really low, and cat kind of spikes a little bit, I feel like Billings just has a little bit more variability than Bozeman.

I: And then you'd say Missoula has the least?

E: Yep.

I: Okay, sounds good. So some dot plots on how many times two different groups of people donated blood.

E: Um, I know we went over this one before but…

I: Well, not exactly this one, it's slightly different, right?

E: It'd contradictory to what I said on the last one, because this one spikes at 3 but I think, I don't know, something about the dot plot, it like making me think because it goes from 1 to 5 versus 2 to 4, then, um, it has a larger variability, I think probably because you're looking at numbers rather than categories, so if it stands for a bigger range numerically, I think that's what makes me think it has a larger range but when you just have categories, it's hard to measure.

I: Right, did it matter, these columns I guess, on these were arranged, if I had done cats first and then dog and then reptiles and rodents? We just kind of pick up the bars and move them.

E: I don't think it would change, if dog was closer to the middle and so it looked more like this, I don't think it would have changed what I thought just because they are all like, you have 4 categories, in each, and so you can't have more than 4 categories, but on this one, you can go, like each of these goes 1 to 5, and this one doesn't, but this one does, and so I think it has a greater variability.

I: Because it actually has the 5 different categories?

E: Yeah.

I: So if it had said all of these, dots in number 3 in group 1, if they had been moved to a different position, just flipped some of them, but still like you know, 2 that have 1 each and 2 that have 2 each and one that has the 6 dots, but if you like rearranged what they were…

E: And this one was still like this?

I: Sure.

E: Um, yeah, I would probably still say that group 1 would have more variability.

I: Okay, and if you like um, hmm. Okay, so basically, just kind of straight based on range is kind of how you're judging it?

E: Uh huh, and well I guess that makes sense, compared to this one, what was I thinking before?

I: So, you said something on this one about if you had to guess, and the person was from Missoula, and you had to guess what kind of pet they had, you'd guess a dog, right? So what if in group 1, if you had to guess how many times somebody had donated blood, what would you say?

E: Three.

I: And how about for group 2, what would you say?

E: I don't know. (laughs)

I: Okay, if you could say 2 things, would you have?

E: 2 and 4, but no, I guess when you phrase it like that, you do have, I guess there is less certainty in this group because you don't know if it's 2 or 4 or something else.

I: So what if you got penalized based on how far off you were, so like, kind of like the price is right, except you always had to be under the actual price, you couldn't go over, but if you were just penalized based on how far off you were, what would be the best guess you could make for group 1 and for group 2?

E: You can't go over?

I: No, you could go over, or go under, but you were just penalized, so like, if you said 2 and then the person actually had 5, the penalty would be like 3 because you were 3 off. If you said 2 and it was actually 1, then that would be a penalty of 1. So what would be like the best sort of?

E: I think the best answer would still be 3, because that's where you'd see the most number of people, and then say you were off, the chances that you are going down like to 2, what is it called? Like there isn't as many people in the 2 or 1 category, there's even less in 1, so there's…

I: So you'd be penalized, if you did get that person with 1 and you guessed 3, then you'd have a penalty of 2, so there would be a bigger penalty.

E: Yeah, but say you guessed 4 over here and it's actually 2, that penalty is 2,

I: Well, would that be? What would be the best guess you could pick on 2? You only get one guess, always have to guess the same thing.

E: Well, if it's always in this range, you might guess 3 and just be penalized the whole time.

I: So you'd have a lot of penalties.

E: But your penalty would just be 1, instead of like 2 a lot.

I: Interesting, so then you're trying to balance the penalties, right?

E: Yeah.

I: Should you, so is it better to often get penalized by 1 or better to do this and sometimes get penalized by 2?

E: Yeah, exactly. I don't know. You have to balance it. I guess it's all this chance, if you pick 2 and it's a large number of 2's, then it's the right choice. But if you pick 2 and then there's a large number of 4's, then it's the wrong choice. Well, and then you keep being penalized by 2, it might be harder at that point, you can't, but it might be smarter to do the 3 and just be penalized 1.

I: So if you didn't know if you'd get this group or that group, if your guess was 3, which group do you think you'd be penalized like the least total amount by? If that makes sense.

E: Group 1.

I: Okay,

E: It's just the proportion that are in 2 and 4 and 1 and 5 are just so much less than 3, that I would feel confident saying that, yeah, is going to 3 show up the most.

I: And sometimes you'd get penalized by 1 and sometimes you'd get penalized by 2,

E: I think the majority would still be in the 3 category.

I: Okay, okay, so, similar but different, so arm spans of two different classes.

E: Oh, this is so hard. I would probably say class 1, kind of the same thing as before,

I: You just went with, doesn't class 1 kind of go with group 2?

E: (laughs) Yes, I don't know.

I: You're kind of flip flopping, what makes this one different? You said that has less variability, group 2 or class 2, what makes this one have more variability, how come? What makes you feel okay about saying that?

E: It's something about the dot plot versus the bar chart, um, I don't know, here it's easy for me to see that these span a different range than this one, but these two both span the same range, um, and in that, in class 2, the majority of, you have 50% that's between 65 and 70, so it looks like this one.

I: But the difference is you have more bars on that end,

E: So it has like the same range, but it's a little bit more uniform than this one, so…

I: So it has more variability? Because it's more uniform?

E: (laughs) It doesn't makes sense.

I: No, there's a reason there are two, that these questions are similar but different.

E: I think that for me, the first sign is range.

I: But on this one, you can't use that.

E: So if these were cut off, or like say, these were cut off, then I would choose this one, because the range is greater than over here, I don't know, it's something about having such a big bar here, and then much smaller, versus all kind of around the same level, or not that, I guess that would make it less variable.

I: Less variable in the arm spans?

E: Yeah, if they're kind of around the same proportion, then there is less variability versus this one. I change my answer.

I: You change your answer, okay you're being consistent. Okay, well, you like that one with the kind of uniform distribution so here's one with a uniform one, they are all exactly the same height.

E: Yeah, um…

I: I'm sorry it's morbid, how old were grandparents when they passed away but…

E: (laughs) Okay, (mumbles) I'd say class 2.

I: Has more variability?

E: Uh huh, what did I say on this one? That I don’t know, I’m going to go on…

I: So at the end on this one you were taking about the heights of the bars and…

E: So these ones aren’t drastic.

I: They’re not super different, but they aren’t exactly the same. But what’s that’s saying about how old the grandparents are and how many are in each age when they die?

E: Well, there’s 20 in each age group, they’re like the same chance, well, in this model, having a grandparent die in 50 to 60, and 90 to 100, so I think there would be less variability in that versus here, you know, how do I explain this, where all the proportions aren’t the same, the frequencies aren’t.

I: So you have more people who are dying between 70 and 80, right?

E: Right, so average, but it still goes out quite a bit on either side. It’s almost like when we were looking at skew, and, you know, let me think of an example, because this is fairly normal, when you have a one that doesn’t have skew and then you have a one that does have skew, there’s more variability because it trails off. And I mean,

I: It’s hard because you don’t exactly…

E: I mean, it doesn’t have skew, but I think that the average number is between 70 and 80 years but then it kind of just like tiers off, so there’s greater variability in the age at which they die, versus a completely uniform.

I: Okay.

E: I don’t know, I’m really bad at explaining.

I: No, no, no, you’re doing good job at explaining, I’m just trying to make sure that… ok, ok, I see what you’re saying, okay, so last one, this will probably bother you because one looks really normal, and one looks kind of normal, right? So.

E: I think I would go with group 1 because it’s kind of that uh, what do you call it? More skew so, it’s weird, because these are like all close.

I: They’re kind of close on heights, but there’s some difference on how high the bars are in group 1, right?

E: Yeah. So, since the majority are between 70 and 80 here, you have 50% alone just in this one category, then 20 and 5 in the other one, so it just kind of like tiers off, so I think there would be greater variability in how much they spent on Halloween costumes. Because even though, you would think, that’s what you’d go to, you’d think most people spent 70 to 80 dollars you have a big range in numbers.

I: But that's the same range of numbers in group 2, right? It still shows their somewhere between 50 and 100.

E: But necessarily, you wouldn't go immediately and say that most spent between 70 and 80 dollars over here, there's only a frequency of 30 here, whereas over here, it's 50.

I: And that is half of 100 people. So if you had to guess what someone would spend in group 2, what would be your best guess if you could only pick one ten dollar range? What would you choose? Group 1 you kind of said you'd definitely pick 70 to 80, but for group 2, would you choose the same or?

E: Right, I would pick 70 to 80, because it does have a bigger frequency, but I'd be less sure of myself because each of these are fairly close, within ten or?

I: So you really wouldn't have been very surprised if someone spent 60 or someone spent 90.

E: No.

I: Okay, but for group 1, would you be more surprised?

E: Yes.

I: Okay. So you'd still say group 1 has more variability?

E: (laughs) I don't…

I: No, that's okay, no, no, that's okay, I guess, is there any relationship between your best sort of guess and how often you'd be wrong with variability? Or like how much you'd be wrong by, since there's no way to guess how many dollars and cents exactly? Does that, go with variability or is it just totally different?

E: No, I think it goes, I think I'm just thinking about this two different ways, and that's why I keep switching back and forth, because I don't know if you look, but how I was explaining with this one but the greatest proportion's between 70 and 80, and then …

I: Sometimes you're still wrong, away from that.

E: But when you phrase it like, uncertainty, I wouldn't be as certain on this one, which maybe indicates that there is more variability, because you don't know exactly how much someone is going to spend on a Halloween costume, whereas on this one, you can be pretty confident.

I: Sure, okay, sounds good.

E: I think I'm just changing my answers.

I: Yeah, I think you're thinking about it in two different ways and struggling between those two ways

E: Yeah.

I: That’s okay, that’s okay.

Emily Interview 3

Interviewer: … questions will be familiar and um, there's the first one.

Emily: Umm, I'll still go with group number 1 has more variability because of the wider range.

I: Okay. Okay. Um, and just kind of range and that's it?

E: Yep.

I: Okay, sounds good. Um, now that you can't use range…

E: (laughs) I think class 2. Um, gosh, it's really hard to describe why but, I think because the frequencies are so varied, that makes me think that there is more variability. Sorry to use the word…

I: No, that's okay, so because… kind of because the bars are different heights, or like more different than in class 1?

E: Yeah.

I: Okay, so the people in these two classes, are like, if you had to kind of describe their height, how are they, I guess, what would a verbal description be of the people in class 1 and what would they look like if you lined them up by height?

E: Um, their range is between 55 and 80 inches. And it's somewhat even throughout the distribution of heights.

I: So if you're lining then up from shortest to tallest, what does that I guess, look like? Does that make sense, what my question is? So like, what I would say is that in class 1, right here 55 to 60 is pretty short, so I'd say there'd be you know, a lot of short people, and then a lot of people kind of in the middle and then a lot of really tall people, because 75 to 80 inches is pretty tall. Umm.

E: Okay.

I: So I guess, is that less different than in class 2, where you have a few short people, a lot of people in the middle, and then a few tall people, tall people I guess.

E: I'm just trying to picture it in my mind.

I: Um, yeah… I know it's hard, I think you want to picture them lined up from shortest to tallest. But how does that, if you could make a graph of the top of their heads, where like, does it jump up or like um, yeah.

E: Well, there is definitely more in the middle here, so you see just a few people on the ends and then like a lot, well not high, you'd see really tall, really short, and then there'd be a lot here in the middle. So when you phrase it like that, it makes me this that this one had more variability.

I: Class 1?

E: Class 1. Because there's more on the ends, further away from the center, the 65-70 inch range and you have fewer in that center range.

I: So there's a difference, right, we were talking right now about their heights, and before you were kind of talking about how many people were in each category.

E: Uh huh.

I: So if we want to know about height? So your final answer I guess is? Just to make sure that I know kind of …

E: I guess, class 1.

I: Okay.

E: That picture made it a little bit different in my head than just looking at a graph.

I: Mmm hmm. So it's more than just looking at the bars, right? Okay, okay. Well, a different situation then.

E: Gosh now I'm just going to be thinking of that. Um, probably group 2. Even though, on the last one like, group 1 was a little more uniform like this one,

I: Uh huh.

E: These ones are little bit closer together so you don't have such a high frequency in the center range, but you still have that um, like the values are different for each interval. Rather than being all the same.

I: Uh huh, so, what does it look like if these people are coming through the finish line and they all started at the same time. What would it look like in group 1 for when they're finishing?

E: 20 people finishing between 50 and 60 seconds, like together, the same thing for 60-70, like 20 runners,

I: So they keep coming through kind of steady-ish?

E: Yeah.

I: Okay.

E: And then for group 2, you'd have like 15 coming in and then 20, so I guess it's less predictable how many are going to come through versus, this one is like a steady rate.

I: You kind of know...

E: How many are going to pass through in that time frame.

I: Uh huh, okay, so group 2, more variability.

E: Yes.

I: Okay, and another one.

E: (laughs) Hmmm. I think I would choose class 2. Um… because this one, kind of going back to the height one, um, you have a lot of your scores are in here, so if you had all these people, I don't know how you'd line them up but…

I: Line them up with a sign saying their exam score, right? (laughs)

E: (laughs) But you know, you'd have a lot in these in these last 3 categories I guess, and then just a couple down here, versus this one your distribution is more spread out.

I: Okay.

E: Um, yeah, I think it's less about the center, because this one is obviously higher than this one, but because there are more people, rather than… sorry.

I: No, no, it's okay.

E: Where we see the most people in these two categories, this is kind of more spread out over each of the values, rather than having like 9 here and then there's at least 10 in each category over here.

I: Okay, okay sounds good, so class 2, more variable?

E: Uh huh. (Whispers) The trick one.

I: I know, I saved this one for last.

E: Um, I think I'm going to go with them same thing, just because the majority of the times in group 1 are between 70 and 80 minutes and then you kind of fall off whereas in this one, there's, like I said before, um, more spread throughout each of the times.

I: Okay, so group 2 would be more variable?

E: Uh huh.

I: Ok. So I know all of these said, “don't make any calculations.” If you could have made some calculations, what sort of things would you want to calculate? (pause) Or none at all? I mean…

E: Standard deviation. I don’t think that the mean is so important, but looking at your spread, since the range is all the same on these except for one, are the same, I think the standard deviation would give you a good idea of, I don’t know, like how far… obviously, this one is going to have a smaller standard deviation because of people in these groups, versus this one, might be a little bit bigger since your frequency is higher in the groups farther away from the center.

I: Okay, sounds good. Any other calculations? Or would that be?

E: None that I can think of.

Megan Interview 1

Interviewer: Here we go, you actually answered these questions before but I kind of wanted to know more about what you're thinking and so here's dot plots, two groups of people and the number of pets they have.

Megan: I kind of remember this one, this one way hard. Um,

I: Was it the dot plot that made it hard or what?

M: Well, I think when I think about variability, I think of more choices and this one is like a little bit more spread out I guess but it doesn't have as many options. So I think I'd go with that one.

I: Group 1.

M: Because it is spread out over a further period even though it has a lot more in the middle, I would go with group 1.

I: Okay, sounds good. And then, some histograms, I kind of see these as very similar to those previous dot plots but certainly they are also different and the situation is different.

M: And in this one, I would say that one too, and I guess, maybe that's the opposite of what I was thinking last time, but this one definitely seems like it was more spread out over the heights.

I: Class 1 is.

M: So it would have more variability.

I: Okay so and this one, you kind of have five bars in each, so if you don't have um, if there isn't kind of a difference in the number of choices that you can see, you can then what are you, how does, what part of it makes it seem spread out?

M: Well, because they are all closer to the same height and this one has some very obvious differences and these ones are like closer, I would say that the results would be more spread out.

I: Okay, so the bars are all kind of similar heights, more so than in class 2.

M: Uh huh.

I: Does the like, position of where the bars are, does that matter at all? I mean, Class 1 you kind of have taller bars on the ends and in class 2, you have a taller bar in the middle?

M: Um, yeah, in like, I mean, not really in this case to me, I guess that is part of it, yeah, these ones are taller but like, this one is not as short as those ones, but then if this one was skewed, like this one over here went down like that, I think that would be more noticeable that it wouldn’t have as much variability because more of them would be at the end.

I: Okay, so if you rearranged the bars in class 2 from tallest to shortest, you’d say that there would be less variability?

M: Uh huh.

I: Okay, don’t worry, there will be some skewed ones eventually. So these are distributions of how long it took two different groups to run 400m, and which one would you say has more variability?

M: I would definitely say that one.

I: Group 1.

M: Because they are like all the same length which means like, this one, you could tell, had an obvious like average, right in the middle, but this one was really varied, there wasn’t anything that stood out a lot, variability…

I: So it had more variability because all the bars were the same height?

M: The bars were the same.

I: Okay, okay, and here, the promised skewed distribution.

M: Okay, yeah, I’d definitely say family 2 because family 1, you can see it definitely has a greater amount of the amounts were way on this side and family 2 was more spread out among all of them.

I: Okay, if you had to estimate like where the median would be on these, what would you, what would you estimate it to be?

M: This one would be right in the middle, family 2, just because, I mean, they are all equal, I guess.

I: So, okay, right in the middle, could you say that it is exactly 75? Or?

M: Well, you couldn’t say that, because they are going by 10’s, but um, if it was like, you’d have, maybe I’m over thinking that, but you’d have to see each individual bar to be able to say that it was exactly in the middle.

I: So you’re saying like each data point, you’d have to know exactly what they are?

M: Yeah, to say it’s exactly in the middle, I think, so right in the middle, so this one would have, we were saying the median right?

I: Yeah.

M: Would be right over here,

I: Okay, so somewhere between 80 and 90 for family 1? Okay, so in family 1, you have some scores that are very far from the median, does that um, does that matter? Does the position of the scores, with the placement of the median, does that change the variability at all or does that not really matter?

M: Um, well, it would make the mean, it would skew the mean, that would make the standard… I think I'm over thinking that, but like the IQR for this one would be actually would be like closer,

I: For family 2?

M: No, it wouldn't. No, it would be farther apart, because there are more people over there and so the IQR would be in family 2, but in family 1 it would be closer to the mean, or the median, because more of them are over there.

I: Between like 70 and 90? Okay.

M: Yeah, and so I think this one would have a bigger IQR and that one has a smaller IQR because it varies more, I think.

I: Okay, and you had kind of mentioned standard deviation, does that, um, what… I guess, it's kind of hard for me to ask you to estimate the standard deviation, right, but, which one do you think would have a bigger standard deviation?

M: Ooh, I think I just realized I was looking at these wrong, no, this one would have a bigger standard deviation.

I: Family 1?

M: Family 1, I think, because this mean is like, the mean would be like over here.

I: So somewhere around 80?

M: Skewed lower, because of those, not really outliers, but those ones would drag the mean lower, but there would be a lot more of the results up here, so I think the standard deviations for family 1 would be bigger than the standard deviation for family 2. Because… Well,

I: Do you think if you picked the data values, you could make it work out either way? That one would have a bigger standard deviation than the other? If you could, we shouldn't make up data, right, but if we did, um, like, you could pick whatever source of values as long as they would fit and still make this histogram, do you think that you could make it so either family have the bigger standard deviation?

M: Um, probably, I guess, I don't know, it's hard to like imagine like a different…

I: Because when you imagine this, you kind of imagine that things are even across,

M: Yeah.

I: That they're kind of you have like 50, 51, or two 50's and 2 51's and 2 53's, somehow to make it fit evenly across, but you do think if it was like?

M: 55 to 60 and 65 to 70, and they would be more like that?

I: I mean, you could… you could certainly, every single value could be 59 and you could have 20 of those, still look like that, you could still make it so that that histogram wouldn't be lying in any way, it maybe isn't giving you as much information.

M: Yeah… (laughs) I guess,

I: So you're not sure?

M: I'm not sure.

I: So it's an interesting question.

M: Yes.

I: Okay, okay, so last one.

M: Ok, I'm going to think about this for a second.

I: That's, you're totally allowed to.

M: Okay. Okay, so I think, I still want to say that class 2 is more varying, because it's more spread out over the same amount of space, but I guess, it's kind of confusing. The more I think about it, the more I trick myself, I think.

I: Ok, so at first glance, you'd say that class 2 has more variability.

M: Uh huh.

I: So on other problems, you talked about the IQR and how they were the same or different, would that method work on this one as well?

M: So here is the median.

I: So around 75-ish.

M: Around 75, I would guess, so the IQR would be somewhere around here, like between 85 and 65, so that would be 20. And so the median on this one would be like here,

I: So like around low 80's, you're thinking?

M: Yeah, probably, so then the IQR for that one would be like… actually, yeah, now that makes me think that class 1 has a bigger variability, so if I was looking at the IQR for this one, it would be more spread out.

I: So somewhere between 65 and 90-ish, or like 60, okay.

M: So maybe that one would have a bigger IQR but still just looking at this one, it makes me feel like it's more varied because looking at the distance, it's still more spread out, which in my mind would make it more varied, but I guess if I was to actually calculate the IQR on class 1 it would actually be um, different.

I: Okay, sounds good.

Megan Interview 2

Interviewer: Have you had a chance to do the moodle questions again?

Megan: Yeah.

I: Okay, so these are probably going to look familiar, as before, but ah, I know it's a little more exciting if you don't do them ahead of time, but that's okay.

M: So, variability, right, that's what we're looking for? So I would say Billings has the most variability. Because, um, it's more spread out between, like in Missoula, you can tell most people own dogs, in Bozeman, more people own dogs and cats, but in Billings it's like a lot more evenly distributed among all four types.

I: Okay, but it doesn't really matter how they were, like so these bars are arranged so it looks kind of skewed or something, or can you even say that in, ah?

M: Well, you could say it's a little skewed but there all a little bit skewed to the right, I guess, but I don't think that that really matters, because there's just more variability in Billings of owning more different types of pets.

I: Okay, okay, sounds good. So then Missoula has the least?

M: Yeah.

I: So, these dot plots.

M: So, originally, these kind of tripped me up, but now I've decided that this one has more variability, group 2.

I: Okay, how come?

M: Because the mean would be um, here, between the two lows,

I: Okay, so it would be 3?

M: So it would be more spread out, away from the mean, I think. Because there's more… well, maybe not on this one. Normally with bimodal, I think of it as the mean is in the middle, so, there's more on that, outside of the mean, so I feel like that's more varied. But with this one, there's only two outside of the mean, so maybe, I still think that that one has more variability

I: Group 2?

M: Yeah.

I: Okay. Yeah, so, um, okay, it doesn't bother you that there's a single 1 and a single 5 in group 1?

M: I mean it does, but I still feel like these ones are more spread out away from the mean because there's more of them outside of the mean.

I: That, aren't right at the mean? Okay, and because they're both symmetric, you can just kind of look and see that the mean should be right in the middle.

M: Yeah, uh huh.

I: Interesting, so you have changed on that one.

M: Yeah, I guess I don't remember what I answered before, which maybe is a good thing.

I: My guess is that you answered it the other way, because almost everyone answers it the other way.

M: That's funny. So on this one, I'm going to do the same, because it's bimodal, so this one is the mean, and this one is the mean, so, I feel like this one is more varied because there is more answers outside of the mean, so this one in my mind would definitely have a bigger standard deviation.

I: Sure, class 1 would and class 2 wouldn't?

M: Yeah, and that makes me feel like it's a lot more varied. Yeah.

I: Okay, so I know you can't calculate the standard deviation in your head, so how did you come to the conclusion that that would be a bigger standard deviation?

M: Okay, so in my mind, standard deviation, I don't know this might be totally wrong…

I: This is a hard question, this is why I'm asking.

M: Standard deviation is like how far the average one is from the mean. And I don't know if that's the actual definition, or anything, but that's how I think of it, how far away most of them are from the mean. In this one, the mean is like right between 65 and 70, and so most of them are a little bit closer to that. In this one, the mean would still be the same, but I mean, more of them are more farther away.

I: So you're thinking about like the distance between the mean and the values?

M: Yeah, like there are more people that are further from the mean.

I: So it's bigger. Yeah, I think that's a really, it's not exactly what the formula says, but that's a good way to think about it, so I think, yeah. Okay, well.

M: Okay, this one I don't like either. I think class 1 would be more varied, and for kind of the same reason, because this one has more that are right around the mean, and this one has a mean in a similar area, but there's more spread out, like it's very spread out.

I: Okay, so, there are more ones that are in the tails?

M: Uh uh, there's more in tails which would kind of skew, not skew but…

I: So, um, in this distribution, do you have enough information, if you were to calculate the standard deviation, could you do it?

M: No.

I: How come?

M: Because you don't know like, the individual values, how people answered each one, I don't think you can calculate it.

I: Okay, so, um, do you think that um, so, right, so we don't have the individual values, so like if we could make up the individual values, do you think you could do it in a way that class 1 would actually have less variability? Like, could you pick them to be certain numbers, and if you picked class 2 to be certain numbers, could you like make it so that you'd be reasonably sure that the standard deviation of class 1 would be smaller and class 2 would be bigger?

M: I don't know, I feel like you couldn't. You could like make these intervals a lot smaller, well…

I: So, you could pick anything, so if you wanted to say, pick them all to be 51, or 59 or you could do two on each value, or something like that, if you had the option to pick?

M: So to make this one, class 1 have the smaller standard deviation?

I: Yeah, smaller standard deviation.

M: I guess maybe you could if you put like, if the mean was still 75 and you put most of them right at the very ends of those .

I: Kind of towards the center.

M: Towards to center, kind of they're still in the 50 to 60 range, right at 59, you could probably make this one have less variability than this one, say if all were 51 and 99, and even further from the mean, I think maybe you could, but it would depend on the interval, maybe.

I: And how many values you can pick, great, that's a hard, it's hard to just look at it, it's something you want to try, just make up the numbers and it could be that way.

M: Yeah.

I: So there could be some. Do you think it would have been possible on that previous question, I mean, this is a lot more dramatic, right? With the arm spans, you have a lot more people in those… tails, I guess. They don't really look like tails.

M: I don't really think so, there's only like 5 in the ends out here, where up here there's almost 30, away from the mean. And maybe it's possible but I think it would be more of a stretch.

I: Yeah, I don't necessarily think it would happen by chance, but, um…

M: I think mostly the shapes would stay pretty similar. I mean, usually you see that, when you shrink the interval, see little changes but you still see the same.

I: Like general, overall shape, sure, sure. So like yeah, in the grand parents one, you still kind of see, that even if you have the bars be half as wide, it still kind of look similar across the top. Yeah, that's true, that does seem to be how normal, just like regular sorts of data that we see are. That's a good point, yeah. So, the last one.

M: Okay, so, I think that group 2 has more variability, again because I mean, it's the same interval, this one is a lot more spread out away from the mean, these look like two things are they? Pretty close, so this one definitely has a lot right around the middle, it definitely has the most between the 70 to 80 range, which is where the mean would be it has like a ton, and doesn't have but they have the same around it. This one has more on the outside, group 2 does, so group 2 I feel like would be a lot more varied, because it's more spread out like…

I: Yeah, no, yeah, yeah. So this one, do you think you could possibly fix the numbers?

M: Like make the standard deviations different? Yeah, I think you could for this one because like they're the same shape, I mean they're not, but they're uni-modal and like symmetric, so this one I think would be a lot easier, if you adjusted the intervals, they might start looking more and more similar and then group 2 might start looking a lot less varied, if [undistinguishable].

I: Sure, yeah,

M: But the way I'm looking at it right now, group 2 seems like would be more varied.

I: Good. And kind of how spread out they are from the mean, that's how you think about variability?

M: Yeah, that's how I think variability, that's what I think about mostly.

I: Sounds good, thanks.

Megan Interview 3

I: Let me know if it doesn't let you go back to do it because maybe I can reset it or something.

M: Okay.

I: Okay, oops, why did I go backwards? Okay, we'll start here.

M: Okay, so… I think group 1 would have more variability because it's more spread out overall, but I think I remember this one and it gave me kind of trouble because this one does have more than are away from the mean, but I still think that group 1, because it's overall more spread out, would have a little bit more variability.

I: Okay, um… and are you thinking, even though it says, "don't make any computations", are you thinking about any computations when you say that or is it just kind of an overall, spread out, I guess?

M: I mean, I guess I always kind of look at it and think about the standard deviation, but that's a really hard thing to think about, you can't just calculate in your head.

I: (laughs) I certainly can't calculate it in my head, I can tell you tell.

M: So, um, that's what makes me wonder about group 2, that there's more than aren't right at the mean, like this one has a lot right at the mean and not as many away, but they're more spread out overall. So that's what kind of trips me up because I don't know which one would have like, a bigger standard deviation without calculating it.

I: Okay, sounds good.

M: Okay, so, class 1, I think, would have more variability. And like again, they're spread out over the same distance, well, maybe not exactly, but… and so class 1 would have more away from the mean.

I: Okay, so, and that one's easier to tell than the last one because?

M: They're kind of just spread out over the same distance and like.

I: Before you kind of had to account for one having a smaller range and the other one having a bigger range?

M: Yeah.

I: Okay, okay.

M: So this one, I would say group 1 again. And kind of the same thinking, they're both spread out over a really similar range at least, but this one's like more uniformly distributed over that whole range, which means like more variability. (laughs)

I: Okay.

M: So it’s more spread out, there’s more results, like, away from the mean.

I: Okay, do you think, if you could pick numbers for these running times, make people come through at different times, that you could make it so that group 2 would actually have a bigger standard deviation?

M: Yeah, you probably could because like um, if these little bars were a different interval, it might change like, how they looked and that would make me change my mind.

I: So like, if I did a bin width of 2?

M: Sure.

I: Okay.

M: Hmmm… (mumbles) Well, I guess I would say class 2 would have more variability, but this one is hard for me. Because they’re both spread out over the same range, or approximately, and this one would have a mean somewhere around here.

I: Like 80-ish for class 1?

M: Yeah, and this one would be 75. Well, maybe it would be class 1, because class 1 has these ones that are way over here and its mean would be more towards the higher side. So maybe that would make the standard deviation kind of bigger and the IQR bigger too.

I: Yeah, so can you um, can you kind of tell where the IQR might be on these two?

M: Um, well I can kind of like estimate, well this one would be right around there, 60 to 90ish.

I: For class 2?

M: Class 2, yeah, and class 1 would be like 95, and this would bring it down lower, I’m pretty sure, because it’s skewed, this was a long time ago…

I: And there are 100 people in each class, that makes it easier to…

M: So then it’d probably be about 55 maybe, so I guess, that’s not the IQR, that’s where it would be between though. But that would make this one have the bigger IQR I think, class 1. So maybe I would go with class 1.

I: Okay, one last question.

M: Wow, that was quick. On this one I would say group 2 has more variability because they’re again over the same range but this one would have more that are away from the mean at least in the way it’s set up, like this.

I: Okay, so, would you, we talked about a few different named distributions, would either of these fit into any of those names?

M: Like uni-modal? Is that what you mean?

I: Sure, or uh, we talked about like a the standard normal distribution and then t-distribution, do you feel like?

M: Oh, yeah, because the thing we talked about they had to be like unimodal and mostly symmetric and both of them are, um, I know that the t-distribution has like taller like, ends, um

I: So, bigger tails?

M: Yeah, so that one would like fit into a t-distribution, but I think that you could use the normal model for both of them, at least approximately.

I: Okay. Cool. Do you have any last thoughts about variability, or, um?

M: Um, I guess the hardest thing for me about all of them was like variability seems really vague, just like, well, what exactly does that mean? That always kind of gave me a little bit of trouble.

I: It seems like you've kind of settled on this, like the distance from the mean, how many data observations are away for the mean.

M: Yeah, I always think about, is it more spread out overall? Is one thing I look at and then also, how far away from the mean.

I: Okay, great, sounds good, thanks!

Tim Interview 1

Interviewer: Okay, here we go, so, you answered some moodle questions. Besides for being willing to be interviewed, your moodle questions were part of how you got selected, so you saw this one before, but looking at two different groups have and the number of pets each one had, which one would you say had more variability?

Tim: See, when I started this, this one made sense, because there is a greater range, from 1 to 5, but now I'm thinking there might be more variability in 2 because it's bimodal. So I really don't know, I'm just going to go with my original guess and say that there is more variability in 1 because it has a greater range.

I: Okay, good, I'm glad that you kind of by the end, felt confused, because I think that variability should be a bit confusing, and there should be some things that are pulling you, I liked that you said that group 2 was bimodal, so that is somewhat different. And when you noticed, were there any measures of variability that you considered, when you thought about the bimodality, so you mentioned the range, and we could compute the range for those two data sets and group 1 would have a larger range.

T: Initially I didn't, but now I see that there is a greater chunk here and a greater chunk here, which tells me that the data is more spread out so that might mean variability in a more holistic sense.

I: Neat, I like it. This is the one about the histograms with the heights, so the question, which class has more variability in their heights, but I think this is kind of interesting, to contrast that group 1 is kind of similar to class 2 and group 2 and class 1, how would you make a decision?

T: I think this one, I was like, well, there's obviously more variability in class 1. Which is totally the opposite thought process I went through on this.

I: Well, there is a difference, the bars are uh, the range appears the same, unlike the last one.

T: But then the question says the variability in their heights, this one obviously has more variability in the heights.

I: In the heights of the class members, uh, yeah…

T: Oh yep, class 1 probably. Class 1 because it's just more numbers spread out over a greater range.

I: Even though the ranges kind of look the same?

T: Yeah, because there is higher frequency. This one seems to have a bunch of frequencies in the middle, and then it tapers off and this one has a lot of frequencies in the ends, and then it kind of drops down to the middle.

I: Okay, sounds good. So, slightly different, right, this one is about running times, but, there are some similarities about the amount of frequencies in the graphs and similar scales, so variability.

T: So this one was getting tough for me, but based on the thought process I've been going through, I'd say there is a generally a higher frequency spread out over times, therefore there is probably more variability in group 1 than group 2 because in the middle, I guess that's the median, that's uh going to have a higher concentration close, and that's more of a standard deviation looking curve, so I'm going to because more of the data is closer to the center that one's going to have more variability.

I: Sounds good, I think your reasoning process works pretty well. How about for families?

T: Yeah, I think I cracked a beer about this time.

I: (laughs) I hope I didn't cause you that much stress.

T: No. No, um, you know, this one was a tough one to look at because this one is consistently more spread out, so using the same reasoning as before, there's probably more variability, because even some of these higher groups like the 80 to 90 range is going to have a greater concentration and the 50 to 60 is going to have the frequency of 1 to 10 which is less. This one, because they are all at 20 all the way across, more variability.

I: In family 2. And um, I guess looking at both of these graphs, would you guess that they would have similar medians or different medians?

T: Probably pretty similar medians, this might have a little bit higher median.

I: Okay, family 1.

T: Because there is going to be more data between 80 and 90 and less basically below 70, so this one is probably going to be a little bit lower, family 2 will be a little bit lower, family 1 a little bit higher.

I: Okay. And the very last one. So this is the one about working and studying.

T: You know, you could just stick with the reasoning, I'm going to say class 2 probably has greater variability because consistently, the bars, the tops, are all closer to each other, where class 1 has more concentration, but then, variability could also be, you know this one only goes up to about 30 in class 1 you know the frequency is going up to 35, 33 let's say, so because there's numbers that exist here that don't exist here, maybe that's variability, and now at the end of this I've realized that maybe I don't know what variability is.

I: Okay, so, what did you think it was in the beginning, if you had to give me a working definition, what would you say variability means to you?

T: I mean, variability is just how much differences there are. Variability is the variety, might be the root word, the difference, it might be synonymous with variety.

I: So thinking, we're looking at graphs of things, so when you're saying differences, I mean, all of these have the same scale on the x-axes, but are you I guess worried about, uh, I don't know how to ask the question, um, I guess, are you, in some of the first pictures, you were kind of saying that, you were explaining where the data was, and how here, in the heights that there was a bimodal one versus one that was kind of unimodal and symmetric, in those ones it seemed like you were kind of saying the position of where most of the data was important, so when you look at this last one, do you see any differences in the positions of most of the data and does that affect the variability?

T: I guess I'm thinking variability, it is a dichotomous thought process which is a stupid trap that stupid westerners fall into but, if it's consistency versus variability, then if I'm shooting an arrow at a target, and the arrow is hitting the same spot like the robin hood thing, each arrow is splitting the next one, that's very consistent shooting. if there's arrows all over the place, then that's a lot of variability on that target. So when I see this, it says there's a lot of consistency, that there are a lot of people close to the mean here, it's not very spread around the target, so that's why I'm thinking consistency and variability here.

I: So for the height example. Ok, and um, certainly these pictures, in the working and studying example are not quite as dramatic as to where the data is different, so um, I guess, does that, let's say on this class 1, if we has moved this 50 to 60, if it had actually been up here, 100 to 110, would that change how much variability you would have in class 1 or would it stay the same? Would it make it more or less or stay the same if you just kind of picked up those 10 data points and say you added 50 to each of them.

T: I guess using the archery example, you know, if I'm hitting like the bull's-eye, that's pretty consistent, and then if I'm getting out and then I'm getting out a little and further into the rings, I'm getting more and more variability in the shots until eventually I'm off the target altogether, so if I'm going to increase either the x or y axis, then you're going to increase that circle you're shooting at on the target, so you're going to decrease consistency and thus increase variability.

I: Okay. Sounds good.

Tim Interview 2

Interviewer: Some distributions of…

Tim: Variability.

I: Variability, I know, it all comes back to it.

T: Well, okay, I'm just… cause you told me I couldn't go and look it up.

I: I know.

T: So I didn't, it's hard not to. So I'm going to think about what I know, and what I know is snow, and with snow science you have spatial variability, so on a slope, you could have a stable snow chunk here, and if you dig your pit there, everything's stable but then you ski 20 feet over there and something changes, so just the more just randomly sampled throughout the snowfield you're taking your snow data, the better off you are. So I'm going to say that has the most variability.

I: Billings.

T: Because there's just a more there's just more, it's just more evenly distributed amongst everything, so I would say that has the most variability.

I: So explain to me a little more how that relates to this idea of the snow…

T: okay, so in avalanche training, what you're going to do it go dig a hole in the snow, you're going to look at the different layers in the snow, and you're going to determine where the weak layers are and what the snow is going to do on top, and that's your slab. So you have four factors for an avalanche, you have to have a weak layer, you have to have slab, you have to have slope, because flat stuff doesn't avalanche, and you have to have trigger, and that could be another snowflake, that could be wind, hopefully it's not you while you're skiing, that's what you want to avoid, you do not want to become the trigger. So you know, what they used to do is a Rouche Block test, because once you identify, you want to do some sort of consistent scientific experiment to the snow to see how much force it's going to take to see when that snow's going to fail and the what we used to do was to chop out a big block of snow and have a second skier jump on that and you'd look and see what happens, and that takes 2 skiers a lot time to dig a hole to get a block, so this professor, Carl Berkland from Montana State University, snow science program geology he has came up with this extended column test, what you do is you create 1 column of snow, 30 cm by 30 cm by 90 cm, or 30 by 90, and you want to go a meter down, so basically, typically, a skiers' not going to affect more than a meter. And so you can tap a lot with your shovel, and that means I only need one person there and I can get this test done a lot quicker. So me and my partner can go spread out and we can each do some of these because what they're saying is maybe on that one slope, there's some rock that you can't see, out crop, that makes the snow a little thinner there, so it's going to be easier to start and avalanche there and they talk about spatial variability being like land mines in the snow so what you want to do is to get the best scientific assessment, you want to have the greatest number of samples in safe zones and I can say I go really good results and my buddy can say I got really bad results 10 feet over, get different results, so the idea of spatial variability to me means that um, you have a bunch of different snow pack conditions and opportunities to hit that land mine and become that trigger anywhere on the slope that you're skiing so you try to get the best scientific assessment when you make your decision of whether you're going to ski or not ski, the best you can do is to get as much data from as much different parts as possible. So I'm thinking…

I: So you're saying that those are like different parts, you're thinking like that's the east part, west part, north part and south part of some slope?

T: Well, that's even more obvious, I'm looking at aspects, I'm been looking at my computer at home and I've got my coffee at 5 in the morning looking at the weather sites on the mountains and you can see which way the snow came which way the wind's been blowing, the straight snow on the windward side, depositing it on the leeward side so the leeward's going the have a slab that could be dangerous and then there's the sun that affects the north is going to get more shade, so it's going to change more slower, the south is going to warm more rapidly east and west, east is going to be more like north, and west is going to be more like south cause of where we're located here and so I'm saying, one I've already done that work and I've already ruled out a bunch of stuff that I'm not even going to go mess with, based on what I can tell from my computer, then you drive up to the trail head and you say, okay, I can look at the trees and see if the rind is forming on one side and if the wind is blowing away, and then you get more information and then you get to this one slope, and you're thinking about skiing, same elevation, same slope angle, same, and at that point now we're talking about spatial variability within those 20 turns I'm about to take, that could also mean that you might have real thin snow and real fat snow and there's more variability there and this is kind of the same trap I talked myself into last time.

I: Okay, well, let's uh, so that's really complicated, and doing things spatially is even more complicated because it's hard to how it relates to each other, and so Missoula, Bozeman, and Billings do have a spatial relationship, but if you're talking to (6:06) people in Missoula and you survey them and then you ask somebody, what kind of pet do you have and there's so many that say dog, so many say cat, so many say reptile, and so many say rodent, the type of pet that they own, the variability, so I think it's a less complicated type of question but it's neat to hear about a, yeah, a situation where that matters.

T: Ok, so I number between 1 and 100 or let's say 98 and 2, has more variability between that number than the variability of 40 and we'll call that 10, so because there's a great variability in percent, there's a more variability in Missoula than Billings.

I: Percent of people that have dogs?

T: Percent of people that have pets in general and type of pet owned right because you have let's say you have the most variability in the type of pet so that means that the variability here for the type of pet owned goes from the number 98 to 2 and that's a greater spread than 40 and 10.

I: Ok, ok, so Missoula would have the more variability?

T: Missoula would have the most and Billings would have the least. I want to change my answer.

I: That’s okay, that’s what I want to know, if you want to change your answer and what affects your view on that.

T: Which group has more variability in the number of times they donated blood? So that means that number of donations, so that one person donated blood one time, and one person donated blood 5 times, so going with that same logic, we’re going to say the difference is between 1 and 1, 2, 3, 4, 5, 6 as opposed to 1, 2, 3, 4, and 0. So it’s a difference of 4 here, and a difference of…

I: Well, there’s actually 5 dots there, right?

T: 1, 2, 3, 4, 5,

I: So from 5 to 0, compared to?

T: 1, 2, 3, 4, 5, 6 to 1 is 5 either way. It’s the same!

I: Do you feel like those have the same variability?

T: Well, if I use that logic, then, yes.

I: Can you use that same logic, I guess is? Here you have bar graphs, here you have dots plots, can you use that same logic?

T: Good point, well, the other way, I could say the variability here, you have people that have done it 2, 3, or 4 times, so there’s only representation in 3 groups, where here there’s representation in 5 groups. So based on that, there’s more variability here because there’s more groups overall that are represented.

I: So if instead, if it hadn’t been 2 and 4, if this bar had been at 1, and this bar had been at 5, would that, would that still just be straight on the number of bars or does it matter um, I guess the range of the bars.

T: You know, if I had to answer, I’d say that the 5 counting rule doesn’t work here unless the answer is the same, but the question here is what group has more variability in the number of times they donated, so unless it’s a trick question, I’m just going to assume it is going to be one or the other. And then I’m going to say 0, if you count 0 as a number, then we’re getting confused, and so if we try to simplify and don’t consider 0 as a number were to spread that out, you’d still only have representation from 3 groups as opposed to having representation from all 5 groups, since this is not an evaluative assessment, I’m pretty comfortable going through that series of logical…

I: And that logic is more what I’m interested in, than evaluating you. Okay, okay, so the one other types of graphical display, right?

T: Arm spans. Yeah, so we have 55 to 80 and 55 to 80, I: Aww, you can't judge on that.

T: So that's the same, so all things are represented essentially, and then this goes up to 20, well, 27 and this goes, well we'll say that's 11, so a difference of minus 11 is 16, where this is a, that's a difference of 5 and 50, a difference of 45, so I'm going to say based on the second logic I used here, class 2 has greater variability because the frequency, there's more spread out frequency, but then with the spatial variability, seeing logic I was using here for more variability, just saying that overall, that…

I: They're more different.

T: They're just more spattered out, you know, because this is more consistent, there's an obvious consistent zone between 65 and 70 here.

I: Uh huh, for class 2.

T: Where this is just kind of more spread out in general, but you know, when I started looking at these, you gave me a clue and said you can't look at that, so I'm going to go with maybe the same logic maybe you talked me into.

I: Bar graphs.

T: Maybe I talked myself into on the bar graphs, and say class 2 has greater variability, more variability.

I: Because of the heights of the bars?

T: Yep.

I: Okay.

T: The difference between the tallest and the smallest.

I: Sure, and uh, and I guess, you had kind of mentioned, you said something approximately and now I wish you would have asked you right away by what you meant about approximately… I guess… you know, in these bar graphs, we didn't really have any spatial relationship with armspan, is there any spatial relationship with those bars, I guess? that question doesn't really make sense. You know, but what, arm spans, go from shorter to longer, like here, versus these,

T: If I was going to put a bunch of people.

I: I could rearrange these, I could put the pets in alphabetical if I felt like it, or do something else. So is there a relationship in that that isn't in here? And does that change anything?

T: Okay, so if I was going to think about that, maybe if I was going to draw a picture, and there was going to be a bunch of people, let's say a similar number of people in each class.

I: And there are, there are 100 in each of these, just to make it easy to look at, right?

T: 100 in each of these, okay, and in this class, I'm going to have you know, 27 people between 55 and 60, bunch of people between 60 and 65, a few between 65 and 70, and there's just, I see a bunch of arm spans all over the class, where this class I see most people are between 65 and 70, and fewer between 60 and 75 and then the 55 to 60 and the 75 to 80, there aren't as many people, so if I was going to you know, line everybody up, there's probably more overall variability in arm spans in this class.

I: Class 1.

T: Because of that. Does that make sense?

I: Yes, because you're saying that there are more people that kind of have different sorts of values?

T: Yeah exactly, more people with different sets of values where this, there's not a whole lot of big tall bar, that's all a pretty narrow range, where all these ranges are all over the place, you have lesser and greater, but overall the scattering here seems like there might be more variability.

I: Ok.

T: But then the other way that I could look at it is that is depends how fine are we going, this is in inches, what if I go down to 100th of an inch or 1000th of an inch, and then eventually, if my measurement progresses to a point where every single person has a different arm span, then the variability is the same. But because this is measured in inches,

I: So then the variability's would be the same? Each arm span is different?

T: Right.

I: So in that case, there wouldn't be any way to judge if one was more than the other? They would just have to be the same if each was different?

T: Yeah variability means that different arm spans, the number of different arm spans,

I: So this is variability in their arm spans, so…

T: Variability in their arm spans, well they're the same, 55 to 80 and 55 to 80. Cause there's, I don't know, I'm totally looked at this like 5 different ways.

I: And you're not quite convinced which way is the best?

T: (reads) More variability in their arm spans.

I: I mean that part is important, that it is in their arm spans. What do you feel like variability means I guess, just in general?

T: Difference.

I: Okay, so the difference. And…

T: There’s more difference in their arm spans. So if I was going to pick 100 people and I said, you 5 people have the exact same arm span, and you 50 have the same arm span, and you 45 have the same arm span, then we have 3 different arm spans represented in the class.

I: Does it matter if they’re all kind of similar, what if your first group had all like 66 inches and the second group all had 68 and the third group all had 70, is that different than if you had the first group had 55 and the second group had 65 and the last group had 75? Would one those two have more variability, does the actual numbers that they’re reporting matter or are they just categories?

T: Absolute… well, I think that those are probably two different ways to look at it, because one is, out of a hundred, how many different unique arm spans do you have and the other is what’s the overall range between 55 and 80? And if it’s between 55 and 80, then they have to be the same because there’s no difference in variability there. And if the difference is in how many individual arm spans are different, if that’s what variability means, I guess, it comes down to that I don’t know what variability means.

I: I think your definition of just like “difference” is a good working definition for it, I don’t think that you’re really misrepresenting it. Which ones are more different? Yeah, okay. So your final answer is that they’re the same?

T: They’re the same.

I: Okay, are these ones also the same?

T: Yes.

I: Grandparent’s age, just based on range. Does it bother you at all that there are more people like here kind of flat across in class 1?

T: Well, when I look at it that way, I’d say this one is more consistent, they’re consistently dying, 20 years, they’re dying young.

I: Some of them are and some of them…

T: Oh, so 20 people die each decade,

I: Hopefully they’re not being killed off for this study.

T: Right. And this one, there’s a different number of people dying in each, so because there’s a different number of people dying in each decade, and this is more consistent, and this is more different,

I: But the question is about the age of death, right? Not about how many die in each decade.

T: Oh right, the variability is the same, because they’re all dying between 50 and 100. So the variability between 50 and 100 is the same as the variability between 50 and 100.

I: Okay. So is this the same as well then?

T: Yeah, the variability, the difference between 50 and 100 is the same as the difference between 50 and 100, so they're the same. The amount of money they spent, so that frequency.

I: Well, that's how many people spent that, yeah, that frequency.

T: Is how many people spent that much money,

I: Right,

T: But that's not what the question is asking, the question is asking how much money they spent what's the variability, so because they're both between 50 and 100, they're the same.

I: And it doesn't bother you that here there aren't as many people spending 60 and here there are more that are spending between 50 and 60.

T: No, because that's not what the question is asking.

I: Okay. So different kind of, another synonym would be range?

T: Sure, range, so then that would go back to that being the same, and that being the same,

I: So all of those histograms, they should be the same? So now this one.

T: You know, I'm going to say that 0 does not count as a number,

I: So 0 people doesn't matter.

T: So no one donated, someone could have.

I: And the point of putting those there is so that you can see that the scale is the same.

T: But the difference between 2 and 4, the range, the difference will be variability in the number of donation there, so I believe was my original final answer, so I'll say class 1 has more variability the number of times they donated blood, and then going back to this one, just so I can be consistent which has the most variability with the type of pet owned, type of pet, dog, cat, rodent, reptile, dog, cat, rodent, reptile, dog, cat, rodent, reptile,

I: Man, you have some of them in each

T: Some in each, so just going to say that all have equal variability, I'm just going to go with that, at least it's consistent.

Tim Interview 3

Interviewer: Well, these will look somewhat new then. Two groups of people and the number of pets that they own.

Tim: So, more variability, I'm using the same logic that I was using at the end of last session, if you recall, which was variability, because there's a range from 1 to 5 pets here and from 2 to 4 pets there, that there's more variability in group 1.

I: Okay, so just on range?

T: Yep.

I: Okay, nothing else?

T: Nothing else.

I: Does anything else matter?

T: Nothing else matters.

I: Nothing else matters. Okay.

T: More variability.

I: Bigger range. Okay.

T: Here, class 1 and class 2, the range is the same, so the variability is the same.

I: Okay, the distributions look really, really different though.

T: They do.

I: Does that bother you at all?

T: Well, we could say that because there's a greater distribution here, we've got to come up with more variability, I'll say class 2,

I: Class 2 has more variability.

T: More variability because they have distribution that has a greater range.

I: So there's a bigger difference in the frequencies?

T: Yeah, like the frequencies, the y-axis here, it looks like you have representation from 5 to 50 and here it looks like you have from 10 to 27.

I: 10 to 27 people?

T: Heights, frequency I guess would be the number of people that height appears.

I: Okay and then in um, okay, so at first one you looked at range, you looked on the x-axis and then we started to look at the y-axis to kind of distinguish in between, is that? If you can't look at x, look at y?

T: Yeah.

I: Okay, okay, and what about these two? People who are running…

T: Okay, well I'm going to go with the x-axis is the same, 50 to 100, but there seems to be no variability there because a frequency of 20 people is pretty much straight across, so that looks pretty boring, no variability there. And there, group 2, the frequency looks like it goes from 15 to 30, so there's definitely more variability in group 2, based on the y-axis because the x-axis is the same.

I: Okay, so if you had these people, kind of like lined them up, these 100 people to run, what would it look like as they're finishing their 400 m, if you could get them all to start? What would it look like in group 1 versus group 2.

T: So group 1, it looks like they're all going to be scattered.

I: Okay, they're going to be…

T: Scattered running times, but it's just 20 people per group, and they're just going to come in in even doses, there's 20, there's another 20, there's another 20 and then this one, you're going to have fewer people that are going to be fast, and then you're going to have a larger group going up to the 70 to 80 second bracket, and then the numbers are going to taper back off again, so they're going to come in in different doses. Right the frequencies at different times.

I: Okay, and that distribution of those frequencies, you'd say, because… um, how do I say that, um, so group 2 is more variable because they're coming in?

T: In different doses.

I: So first you have 15 people trickle through between 50 and 60 seconds, then you have more, then you have a lot of people between 70 and 80, okay. So that would be more variable.

T: Yeah.

I: Okay,

T: More variability in the exam scores. Well, here we go, so we've got 50 to 100, 50 to 100, so that's the same, and here we have a frequency of you know, what's that, 8, and then what? 38, and this one goes from what, 15 to 30, so because you know these students are all getting between, because its between 15 and 30 students difference between the lowest score and the most frequent score, the least frequent and the most frequent scores, here you have um, just, a wider range of frequencies over the various exam scores. Everyone still scores from 50 to 100, I: Uh uh, okay, so you first kind of computed range, what about are there other measures that can be computed…

T: Oh, I'm sure there's plenty.

I: … in a histogram?

T: But what they are, I have no idea.

I: Okay, that's okay

T: I'm not afraid to admit there's some limitations in my understanding.

I: Oh, you don't have to know everything, that's okay, no, I was just wondering.

T: Without making any calculations, which has more variability.

I: And then I ask you what calculations could you make, that's kind of mean.

T: That's right, cruel and unusual. So number of minutes spent exercising. So again, the x-axis, the number of minutes spent exercising has the same range, from 50 to 100 minutes but there's more variability in group 1 because you know some people spent 5 minutes while some people spent 50 minutes, while group 2 is not as lazy or as ambitious, in that some people are going to spend a minimum of 15, but no one's going to spend more than 30 minutes exercising.

I: So, these histograms, is it you had said that they had spent like 15 minutes exercising, where I see that the exercise times are between 50 and 100.

T: Oh! That's right, you're right, I said that wrong, but the frequency, still spend the same amount of time but overall, there's a narrower range of frequency in group 2 than group 1.

I: Okay, okay, so if you had to say, pick a name for the group 1 distribution, if you had to approximate it with one of the distributions we've been using in class, what name would you give it?

T: I would call them both Normal.

I: They both look Normal?

T: They both look pretty Normal, they just look, just differing in degrees.

I: Degrees?

T: Frequency, they both differ in frequency.

I: How many, okay…

T: As that is defined as between the lowest frequency and the highest frequency, group 1 has a greater…

I: There’s more people in the middle, huh,

T: But they both follow the same sort of normal bell curve.

I: Okay.

T: That’s just a steeper one.

I: And the less steep one, okay. Sounds good. I think that’s all I have. Do you have any last words about variability that you want to share?

T: I don’t like variability right now.