UNIVERSITY OF CALIFORNIA

Santa Barbara

# Two-Year College Mathematics Instructors' Conceptions of Variation

A Dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Education

by

Monica Graciela Gandhini Dabos

Committee in charge:

Professor Mary Betsy Brenner, Co-Chair

Dr. Dawn Homes, Lecturer, Co-Chair

Professor Yukari Okamoto

December 2011

The dissertation of Monica Graciela Gandhini Dabos is approved.

_____
Dr. Yukari Okamoto

_____
Dr. Mary Betsy Brenner, Committee Co-Chair

_____
Dr. Dawn Holmes, Committee Co-Chair

September 2011

Two-Year College Mathematics Instructors' Conceptions of Variations



Monica Graciela Gandhini Dabos

# ACKNOWLEDGEMENTS

The writing of this dissertation has been one of the most significant academic challenges I have ever had to face. It has been a long educational journey that is ending with the final copy of this dissertation. I would never have been able to finish this dissertation without the guidance of my committee members, the professors who volunteered their input, and the incredible help from friends and family.

I offer my deepest gratitude to Dr. Betsy Brenner, whose commitment to help me succeed was evident though all the stages of this dissertation. Her feedback and insightful suggestions helped immensely. I want to especially thank Dr. Dawn Holmes for always encouraging me to go a little further and to never give up. I want to also thank Dr. Yukari Okamoto for helping me achieve higher standards of writing.

I want to particularly thank all the instructors that took part in this study: thank you for making space in your busy schedules to become intricately involved in this research.

I would not be writing this acknowledgment page on the final stages of my dissertation if not for the vision of Dr. Howard Resnick, my spiritual guide, who encouraged me to go back to school and pursue a degree. He saw the potential in me when I did not see it. Thank you, Srila Acaryadeva, for your vision and encouragement.

I owe my deepest gratitude to Richard Stocker, without whose help this dissertation would have been impossible: thank you, Richard, for your love, dedication and support in spite of it all. You are the only one who stood by my side from day one to this day, seven years later (about 2,555 days). Thank you - I feel eternally indebted to you.

Thank you, Dr. Larry Kugler, whose insightful discussion guided me to think like a researcher. I want to thank Melissa Zaragoza for her incredible help when I most needed it: thank you, Melissa, for helping me clarify my thoughts and for constantly encouraging me whenever I felt that it was impossible to continue. I could not have done it without you. Thank you, Elea Bayley, for your kind help and support. Thank you, my friend Meveves Cabeza, whose cheerful energy and tasty meals kept me going. To my friend, Jenny Fernandez, who ran along with me for a long stretch just to keep me company: thank you. To the company and valuable help of Melina Allahverdian, who also ran along with me for a good while.

Thank you to my spiritual brothers and sisters who kept me centered through it all. In particular, endless thanks to Barbara Dunaway and Dr. Fred Gamble, who offered me a sanctuary, transforming my life.

I thank my sister, Cristina Dabos, and my brother, Julio Dabos, and his family, for encouraging me all the way from Argentina, following my steps closely and cheering me all the way. To my brother, Gustavo Dabos, and his family: a special thank you for supporting me in more than one way to make this possible.

There are so many friends and family to thank that I would like to write this section as a three-dimensional form so that I can thank and write the names of every single person who made this possible. I want to thank all my friends and family members who are not mentioned in here, but who know that their support helped me get where I am right now. Thank you.

I dedicate this dissertation to my deceased mother and father, Elena Greco and Cirilo Dabos: they believed that education was the only inheritance they could give me. I am sure that they would be very proud of this achievement, which is the fruit of many of their endeavors. Thank you Mom! Thank you Dad!

VITA OF MONICA GRACIELA GANDHINI DABOS
DECEMBER 2011

EDUCATION

Bachelor of Science in Elementary Education, Florida State University, December 1999 (cum laude)

Master of Arts in Statistics, University of California, Santa Barbara, June 2003

Bachelor of Arts in Mathematics, University of California, Santa Barbara, July 2004

Master of Arts in Education, University of California, Santa Barbara, January 2010

Doctor of Philosophy in Education, University of California, Santa Barbara, September 2011

PROFESSIONAL EMPLOYMENT

2000-01: Spanish Teacher, New Academy Elementary School, Santa Barbara

2000-03: Teaching Assistant, Department of Statistics, University of California, Santa Barbara

2002-Present: Math Tutor, Santa Barbara City College

2003-Present: Co-coordinator, College Achievement Program (CAP), Santa Barbara City College

2003-Present: Adjunct Instructor, Santa Barbara City College

PRESENTATIONS

"Motivating students with interactive statistics," Workshop at the California Mathematics Council Community Colleges, San Diego, 2004

"Preparing tutors to assist students in statistics," Santa Barbara City College, 2006.

"Focusing on the understanding of p-value, Type I and Type II errors," Santa Barbara City College, 2008

"Mathematics instructors' conceptions of variation at two-year colleges," Poster presentation at United States Conference on Teaching Statistics, Columbus, OH, 2009

"PCK: The case of statistics," Workshop at the California Mathematics Council Community Colleges, San Diego, 2010

AWARDS

Certificate of Recognition, Santa Barbara City College, 2004

Faculty of the Year, Transfer Achievement Program, Santa Barbara City College, 2009

FIELDS OF STUDY

Major Field: Statistics Education

ABSTRACT

Two-Year College Mathematics Instructors' Conceptions of Variation

by

Monica Graciela Gandhini Dabos

Statistics education researchers are urging teachers of statistics to help students develop a more sophisticated understanding of variation, since variation is the core of statistics. However, little research has been done into the conceptions of variation held by instructors of statistics. This is of particular importance at the community college level because there was a 60% increase in enrollment in introductory statistics courses in a recent five-year period. Moreover, at the community college level only 2% of full-time instructors and 2% of part-time instructors have a degree in statistics. This exploratory study was designed to map the conceptions of variation held by two-year college mathematics instructors. A total of 52 instructors from 33 different California community colleges responded to a survey designed to reveal instructors' conception of variation. All of the instructors had a degree in mathematics – seven of them also had a degree in statistics. Instructors had varied statistics teaching experience: 23 of them had never taught statistics and 29 had taught statistics often. The results of the study indicated that there was a difference in the type of responses; however, their educational background or their statistics

teaching experience did not highlight the difference. The results indicate that some instructors readily acknowledge variability and others do not. A tendency to focus only on the center of the distribution seemed to prevail, and very few instructors gave explanations integrating different aspects of the distribution. The majority focused on the center or on the range. Another salient characteristic of the results of this study was instructors' lack of consideration for context when making decisions about variability in the data. Moreover, this study revealed that instructors were not consistent throughout the survey items. In some cases, instructors predicted variability, but their justifications lacked the appropriate reasoning to support their predictions. This study has opened the gate and laid the groundwork for understanding conceptions of variation held by two-year college instructors. The results indicate that more in-depth investigation needs to take place if the goal is for students to develop a sophisticated conception of variation.

TABLE OF CONTENTS

## List of Tables

# Chapter I

The purpose of this study is to research conceptions of variation held by mathematics instructors at two-year colleges. This chapter is divided into two main sections that provide the grounds for this study: In the first section, the rationale for studying two-year colleges is provided by stressing the importance of statistics, its increasing demand, and the need for qualified teachers. In the second section, the rationale for studying the concept of variation, its relevance in statistics, as well as the complexity associated with the topic of variation is discussed. The combination of these two main sections reveals the significance of investigating two-year college instructors' conceptions of variation. The chapter ends with the objectives of the study, and the description of the structure of the paper.

## Teachers' Knowledge and Two-Year Colleges

The role of statistics education is currently undergoing a profound shift. Globalization and technology are forcing researchers to question the relevance of different statistics topics in a rapidly changing society. The democratization of mathematics and statistics has made these subjects available to a very large, diverse and less specialized student population. In fact, the demand for statistics instruction at all levels of education has inundated the educational system in American schools.

In the ten year period ending in Fall 2000, the number of statistics departments in the United States grew by 68%, while student enrollment climbed 37%, according to a national survey of undergraduate mathematical and statistical sciences in the nation's universities and colleges, both four-year and two-year, organized by the Conference Board of the Mathematical Sciences (CBMS) (Lutzer, Maxwell, & Rodi, 2002). Building on this base, in the following five years, from 2000 to 2005, statistics enrollment at two-year colleges increased by 60% (Kirkman, Lutzer, Maxwell, & Rodi, 2007).

Even though the increase in student enrolment is substantial, there is little information about who is preparing those students to succeed in statistics. According to the CBMS report, 16% of full time and 2% of part-time of mathematics instructors at two-year colleges in the United States have a doctorate degree. In addition, only 2% of full-time and 2% of part-time instructors teaching at two-year colleges have degrees in statistics. These figures are not likely to change any time soon, particularly in California, since those with a bachelor's and a master's degree in only statistics can not teach statistics at two-year colleges, as these colleges require a degree in mathematics. This is of great concern considering that California's is the largest college system in the world, as per the California Community College Chancellor's office home page. Therefore, it is not clear where instructors teaching statistics acquire the knowledge necessary to teach statistics. It could be said that it is not necessarily from the background degree.

Besides not having background training in the subject, instructors face an additional impediment in that introductory statistics courses have been labeled the most challenging courses in the undergraduate curriculum for students (Delucchi, 2007). Also, Yilmaz (1996) states "statistics is a difficult subject for non-specialists, not just from the viewpoint of the student but from the teacher as well" (p. 2). Taking into consideration the lack of background training together with the challenge that teaching introductory statistics poses, and the difficulty that non-specialists face, it comes as no surprise that many of those teaching statistics consider it an unrewarding experience (Garfield, Hogg, Schau, & Whittinghill, 2002).

Empirical studies have extensively, but not exhaustively, investigated students' difficulties in statistics at all levels of the education system (Batanero, Godino, Vallecillos, Green, & Holmes, 2000; Garfield & Ahlgren, 1988; Lavigne & Glaser, 2001; Quilici & Mayer, 1996; Meletiou-Mavrotheris & Lee, 2005). Students' difficulties can be attributed to many factors, including lack of interest in the subject, students' own beliefs about statistics, previous mathematics experiences, and the learning environment (Gal & Ginsburg, 1994; Tremblay, Gardner & Heipel, 2000). Many students experience difficulties because of their limited understanding in various areas of statistics. For example, limited understanding of the arithmetic mean leads students to have problems interpreting arithmetic mean under different circumstances (Watson, 2007). Limited understanding of probability affects students' ability to comprehend randomness or variability (Reading & Shaughnessy,

2004). Limited understanding of histograms leads students to wrong conclusions (Meletiou-Mavrotheris & Lee, 2005). Students' limited understanding also affects their ability to correctly choose what statistical test they need to perform to answer a particular empirical inquiry (Gardner & Hudson, 1999).

These difficulties have been identified as the result of researchers' endeavors to map out students' struggles. One of the goals of this mapping is to facilitate those teaching statistics to understand students' difficulties. Therefore, in almost every case, the researchers at the end of their investigation provided a guide for those teaching statistics. However, "teachers have the same difficulties with statistics concepts as the students they teach" (Shaughnessy, 2007, p. 1,000). Moreover, the results of these studies on students' difficulties not only provide practical applications to those teaching the subject, but they also demonstrate the central role that instructors play in minimizing students' difficulties and in fostering students' understanding. However, there has not been a large emphasis on the part of researchers in trying to understand those who teach statistics, especially at two-year colleges. If the goal is to minimize students' misunderstanding and maximize students' statistical literacy, then there is an urgent need to seriously investigate the statistical knowledge of those teaching introductory statistics courses at two-year colleges. Since the investigation of teachers' knowledge or the knowledge necessary for teaching statistics is rare, more research on teachers' conceptions of statistics is necessary.

**Importance of Variation**

Variation has long been recognized as the essence of statistics (Moore, 1990), and in fact statistics has been viewed as "the science of variation" (Wild & Pfannkuch, 1998, p. 6). Scholars acknowledge that the speech delivered by Mike Shaughnessy (1997) formed the impetus for the recent interest in research about variation as a central theme of statistics education. However, understanding variation and learning how to deal with it are not simple tasks. The complexities of variation were clearly exposed in the seminal article written by Wild and Pfannkuch in 1999: "Statistical thinking in empirical enquiry". This article was written as a synthesis of their previous research done through a series of in-depth interviews conducted to investigate the thinking process of experienced statistical practitioners and tertiary statistics students as they solved statistics problems (Pfannkuch & Wild, 2004).

The authors stated that there are three main messages about variation that everyone needs to understand:

1) variation is omnipresent;
2) variation can have serious practical consequences, and
3) statistics gives us a means of understanding a variation beset world (p. 235).

Wild and Pfannkuch explained that variation is ubiquitous, and indeed it is easy to note that no two manufactured objects are equal; no two organisms are identical or react the same way. Therefore, when data are collected, extra variation is added to this fundamental underlying variation. With this graph (Figure 1.1) the authors

presented some of the sources of variation, illustrating the complexity of understanding variation. To further explain how statisticians deal with variation, the researchers provided the graph below (Figure 1.2) and explored three main courses of action.

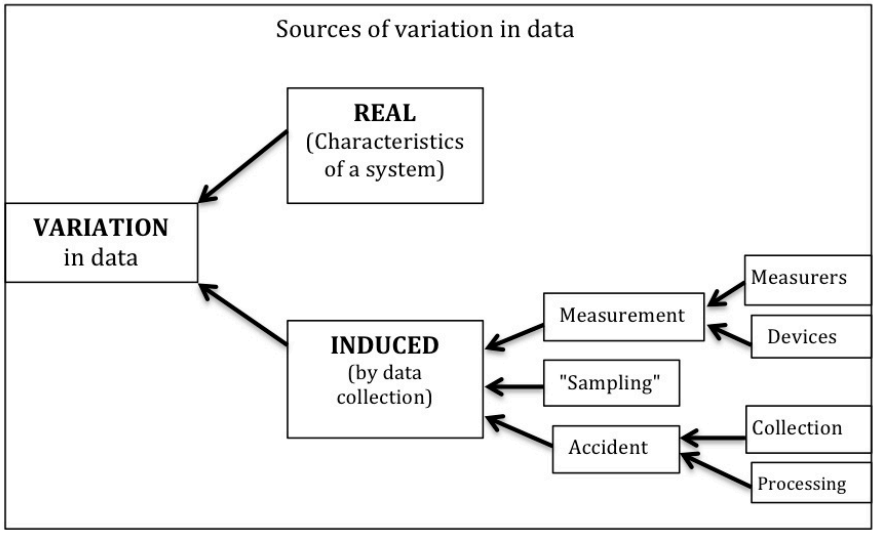



*Figure 1.1.* Practical Responses to Variation. Adapted from "Statistical thinking in empirical enquiry," by C. J. Wild and M. Pfannkuch, 1999, *International Statistical Review*, 67, p. 236.

Firstly, variation can be ignored, for example they explained that one can "behave as though every object or organism is the same or differs in some deterministically known way" (p. 236). They indicated that this approach may be useful in some cases. Secondly, it can be worked around by predicting ways in which a product may be subject to variation and making the product "robust" to that variation. Thirdly, manipulating causes or applying external treatment, for example calibrating a machine, can change the pattern of variation. Statisticians look for

sources of variability by looking for patterns and trying to understand the relationship between variables “Statisticians model variation for the purpose of predictions, explanations, or control” (p. 236). Wild and Pfannkuch explained that even though variation may sometimes be perceived as an impediment to understanding the effects of a particular system, examining the variation may lead to the discovery of statistical patterns.

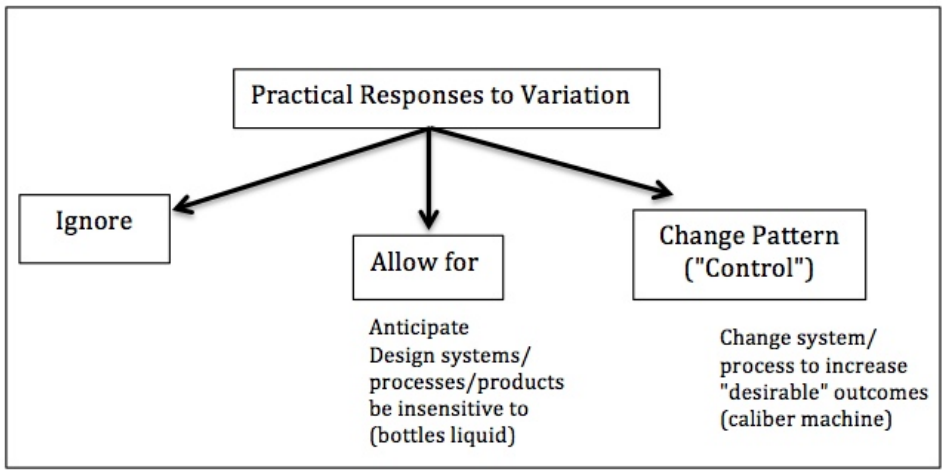



*Figure 1.2.* Sources of Variation. Adapted from "Statistical thinking in empirical enquiry," by C. J. Wild and M. Pfannkuch, 1999, *International Statistical Review*, 67, p. 235.

Figure 1.3 depicts the many ways in which paying attention to regularities or irregularities found in variation can help estimate the underlying causes of variation in a system. The diagram and explanation provide a clear picture of how variation enters into data collection and how to deal with it. However, this is not a simple task - as the authors stated “we are dealing with a complex and sophisticated process” (p. 246). Wild and Pfannkuch explained that variation influences the thinking of every

aspect of an empirical enquiry and it needs to be understood taking the context into consideration.

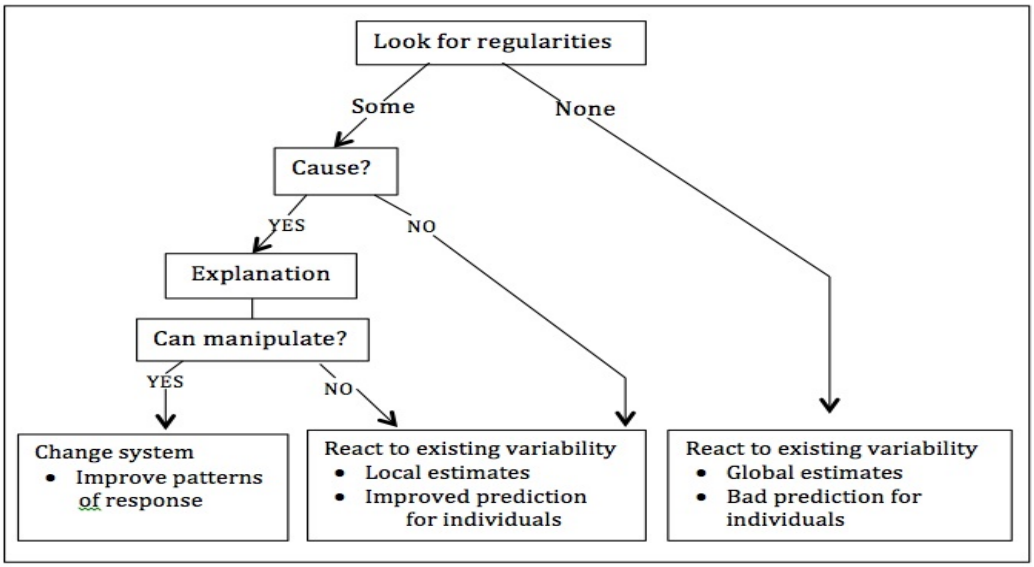



*Figure 1. 3.* Dealing With Variation. Adapted from "Statistical thinking in empirical enquiry," by C. J. Wild and M. Pfannkuch, 1999*, International Statistical Review*, 67, p. 237.

Variation as described above is central to the understanding of statistics, is very complex and at the same time students need to master the concept of variation. Therefore, this should be one of the first areas of investigation of instructors' knowledge. Learning about the conceptions of variation held by mathematics instructors could provide the basis to understanding their ability to deal with variation, as well as their ability to teach it properly.

To summarize, the importance of statistics and its increasing demand have prompted researchers to try to understand students' struggles. However, the knowledge of those teaching introductory statistics has been largely ignored.

Therefore, this project is of prime importance and a step forward in this area of research. Understanding of variation has been recognized as the core of statistics; therefore, the concentration on this topic seems a necessary starting point in the investigation of two-year college instructors' knowledge.

The aim of this study then is to investigate conceptions of variation held by mathematics instructors at two-year colleges by describing their responses to several statistics questions. These questions aim to explore statistics concepts and ideas used by instructors to answer them. In particular, the study focus on their responses to variation in four main contexts:

1) repeated samples, including sampling distribution,

2) graphical representations of two data sets,

3) graphical representations of one data set, and

4) average.

This study also aims to compare the responses provided by instructors according to their background training (i.e. mathematics degree vs. statistics degree) and their statistics teaching experience (whether they have taught statistics or not). There are three main research questions and each question has sub-questions.

Research Question One (RQ1):

What are instructors' predictions in repeated sample experiments?

- Is sample variability reflected in instructors' predictions?

- Are instructors predicting an appropriate amount of variability?

- What statistical concepts do instructors focus on when dealing with empirical sampling distributions?

- How are teaching experience and education related to instructors' predictions in repeated samples?

Research Question Two (RQ2):

What statistical concepts do instructors utilize to describe and/or decide the best way to represent one data set and does the context of the problem guide those decisions?

- in the presence of unusual variation?

- in the presence of several graphical displays?

- in a histogram?

- How are teaching experience and education related to instructors' responses?

Research Question Three (RQ3):

What aspects of the graphical representations do instructors pay attention to in the presence of two data sets:

- To decide which graph has more variability?

- To describe and compare the graphical representations?

- To aid them in making informal inference or decisions?

- How are teaching experience and education related to instructors' responses to graphical representations of two data sets?

The next six chapters will show the process followed to answer these research questions. Chapter two includes a description of prior research studies that investigated conceptions of variation. Three main statistics contexts have been highlighted: Variation in distributions, variation in repeated measures, and variation in a data set exploring the mean. Chapter three provides the methodology and rationale for selecting and analyzing the data. Chapter four presents the findings corresponding to Research Question One. Chapter five presents the findings corresponding to Research Question Two. Chapter six presents the findings corresponding to Research Question Three. Chapter seven contains a discussion corresponding to each research question, and general findings, limitations and implications.

# Chapter II

## Literature Review

This chapter reviews the literature on how statistics instructors and their students understand the important concept of variation. Variation is one of the most important concepts in statistics (Moore, 1997; Watson & Kelly, 2002), with some indicating that variation is "the very heart of the statistics we teach" (Wild & Pfannkuch, 1999, p. 241). As stated by Reading and Shaughnessy (2004), "more research on reasoning about variation needs to be undertaken to assist educators to better equip future students in measuring and modeling variability as they reason about variation" (p. 204).

There seems to be an implicit understanding behind Shaughnessy's recommendation that those teaching statistics possess the appropriate understanding of variation and therefore they need only to understand its importance, to be aware of how students think and reason about variation, in order to be able to teach it to students correctly. However, the state of those teaching statistics does not seem to support the premise that instructors understand the subject matter well. Shaughnessy (2007), making reference to K-12 teachers, indicated, that "teachers have the same difficulties with statistical concepts as the students they teach" (p. 1,000). Moreover, statistics graduate students (TAs) investigated by Noll (2007) showed similar results: "these TAs experienced some of the same key difficulties as middle, secondary, and tertiary students" (p. 335). There are only a few studies (Dabos, 2009; Garfield,

delMas & Chance, 2007; Thompson, Liu & Saldanha, 2007) that provide a glimpse of the conception of variation held by those teaching statistics at the college level. These studies revealed that college-level instructors have gaps in their understanding of variation. Overall, researchers have found gaps in the conceptions of variation held by those teaching statistics at all levels of education.

Although the purpose of this study is to investigate conceptions of variation held by those teaching statistics at the college level, the scarcity of such research in the literature limits the ability to fully reveal the abilities and limitations of those teaching statistics at the college level.

Thus, much of this literature review will focus on studies that investigated conceptions of variation of students at the college level, as well as studies dealing with K-12 students, since the topic of variation has been investigated more extensively at this level. The aim of this literature review is to reveal what researchers have discovered about the understanding of variation at all levels of education. If a clear picture of students' understanding of variation is developed, then a link could be made from students' misconceptions to those instructing them.

The literature review begins with a focus on the few studies that reveal important information about those teaching statistics conceptions of variation at the college level as well as teachers at the K-12 level. It then shifts the focus to studies that investigate students' conceptions of variation.

### Statistics Teachers' Conceptions of Variation

Garfield et al. (2007) utilized the Japanese Lesson Study (JLS) methodology to work with a group of six to eight college instructors of statistics to design a lesson, then teach it, modify it, teach it again, and evaluate students' understanding of variability. The instructors had a mixed level of teaching experience that included novices and experts. Besides the promising results that students' understanding of variability improved after the lesson was modified, the most relevant aspect of this study is that instructors themselves encountered their own misconceptions of variation in the process of developing the lessons. Garfield et al. stated, "through our discussions, our own understanding of variability deepened and improved, even though some of us have taught this topic for more than 20 years" (p. 128). Although the time demands of using JLS are great, the JLS method could be a promising method to improve statistics instructors' own understanding of variability and other statistical concepts. This in turn could greatly improve students' conceptual understanding of variability.

Others researchers have had similar realizations that those teaching statistics may have the same difficulties as students do (Shaughnessy, 2007; Noll, 2007). A relevant study was conducted by Thompson, Liu & Saldanha (2007) who initially began working with high school teachers to improve Advanced Placement statistics courses. However, after realizing that the difficulties that students demonstrated with inference were rooted in their teachers' understanding of statistics concepts of variation and inference, they decided to investigate the teachers. They conducted a

twelve-day summer seminar where teachers got paid half a month's salary. There were a total of eight participants in the study, all of them teachers who had taught statistics or were preparing to do so.

Thompson et al. (2007) conducted three interviews with these eight teachers: one before the seminar, one at the end of the first week, and one at the end of the second week of the seminar. The pre-seminar interview revealed that teachers had the same difficulties as students. In particular, teachers were "predisposed to think in terms of individual samples not in terms of collections of samples" (p. 218). The first week of the seminar was dedicated to different aspects of understanding and teaching inference. The second interview intended to see if teachers had internalized the logic of hypothesis testing after the first week of the seminar. However, the second interview exposed teachers' limited understanding of hypothesis testing logic. The authors suggested that even though the teachers had some understanding of the logic of hypothesis testing, they did not understand its functionality. "Most of the teachers did not understand the logic of hypothesis testing, or if they understood it they thought it was irrelevant to settle competing claims about a population parameter" (p. 224).

According to Thompson et al. (2007), the cause of teachers' difficulties may have been due to the compartmentalized knowledge of probability and statistical inference. Their concept of probability was not grounded on the concept of distribution and variability. The authors inferred that those who have developed distributional reasoning in probability would be better positioned to help students

reason about statistical claims. The other finding revealed that teachers' logic of argumentation was wrong; for example, they believed that rejecting the null hypothesis implies proving it wrong. Their conclusion is that teachers' difficulties with hypothesis testing resembled those of high school students.

Thompson et al. (2007) concluded, "not only must teachers understand students difficulties and the way they might overcome them, but they must adjust their own understanding to support logic of argumentation that is alien to them" (p. 226). This study revealed not only the limited understanding of teachers' statistics conceptions of variation and inference, but also that the seminar did not seem to improve their understanding. It may require more than 12 days to change instructors' long held beliefs and misconceptions. If this study is taken as a representation of the status of those teaching statistics across the nation, there appears to be serious work ahead to improve the understanding of those teaching statistics. Since these eight instructors cannot represent the entire population, more studies are needed for obtaining a clearer picture of the understanding of variation held by instructors of statistics.

Consistent with the previously mentioned studies, the research conducted by Dabos (2009) of six college mathematics instructors' conception of variation also revealed that instructors had limited conceptions of variation. In this study, six two-year college instructors answered a two-part survey that included tasks to assess their conception of variation. The instructors had a variety of backgrounds and experience

teaching statistics, including some who had taught it every quarter for several years, those who had taught it a few times, and those who had never taught statistics.

Dabos' (2009) study was mostly exploratory with the goal of identifying intuitions and conceptions of variation instructors had when dealing with repeated measures, sampling distributions, histogram interpretations and informal inference. The responses varied very little; some instructors integrated a few aspects of the distribution in their descriptions, but most of the instructors tended to focus on the center of a distribution or on extreme values when looking at distributions. Instructors seemed to have compartmentalized knowledge and could not use all three aspects of variability (center, spread, and shape) simultaneously to assess information or draw conclusions. They also showed that their level of certainty on their responses was very weak. In this study, certainty was not formally measured, but was deduced by long periods of silence or by their statements of “I don’t really know”.

Overall, instructors seemed to struggle with ideas of repeated samples. They tended to not acknowledge variation in outcomes of small trials, and mostly focused on theoretical probability and expected values when giving predictions of an empirical experiment. Again, the small sample size of this study makes it difficult to generalize the results to all instructors, but the glimpse showed that college instructors, regardless of experience and educational preparation, seemed to have limitations in their knowledge about variation.

These kinds of studies, investigating or discussing those teaching statistics, presented above, are very rare and therefore it is difficult to develop a broad understanding of the conception of variation held by those teaching statistics. However, these studies resonated a common theme, namely that those teaching statistics experienced difficulties with variation that resembled students' difficulties. The next section of this review covers the research on students' misunderstandings.

## Students' Conceptions of Variation

While studies about instructors are scarce, the studies dealing with students' reasoning about variation provide a deeper picture of the difficulties students face while trying to deal with variation across many statistical processes. This section of the research literature is divided into three main themes: 1) Variation – distributions and their relationship, 2) Average – connection to variation, and 3) Variation and inference – the hidden background needed.

### Variation – distributions and their relationship

Jane Watson (2009) described the meaning of a distribution starting with the typical dictionary definition that describes a distribution in terms of variation. She recognized that the simple dictionary definition "to spread out" may be a good starting point for students. She identified Moore and McCabe's (1993) definition as more advanced since it included the idea of variation. According to Moore and McCabe (1993), "The pattern of variation of a variable is called its distribution. The

distribution records the numerical values of the variable and how often each value occurs" (p. 6). To display such information, the use of graphical representations is desirable.

Watson (2009) also stated "the relationship of variation to the statistical concept of distribution is close, but intuitively variation is a term covering all sorts of observed changes in a phenomena whereas distribution is a more formal notion based on graphs" (p. 34). In the research literature there are two forms of assessing understanding of distributions in graphical representations as a medium to understanding conceptions of variation held by students. One is to provide students with graphical representations and assess where students put their attention when describing them. The other is to ask students to create graphs based on experiments or manipulate graphs using computer programs. In the following section a selection of these studies is presented.

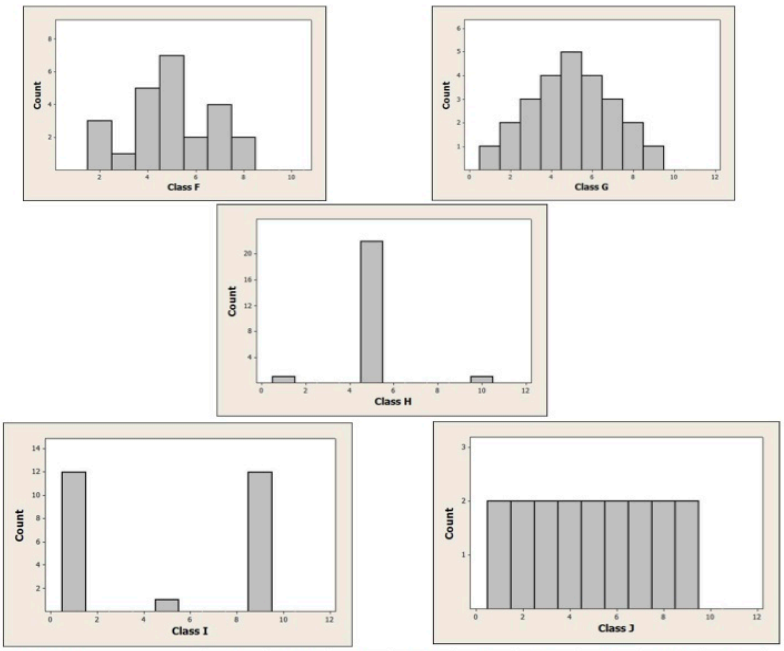
### *Graphical representation of a distribution*

A relevant study conducted by Meletiou-Mavrotheris and Lee (2005) was designed to assess college students' understanding of variation by looking at histograms. Students were presented with five different histograms (Figure 2.1) and were then asked a series of questions; the first asked students to identify which histogram had more variability. The researchers analyzed responses for 162 college students and found that even after their own instruction, their students still had difficulties explaining the variation found in histograms; 70% of the students gave

the wrong interpretation of variation after instruction (p. 4). Also, they noticed that when students were asked to describe variability, they tended to focus on the variability they saw in the y-axis of the histogram instead of focusing on the x-axis and paying attention to how the data varied from the center.

Meletiou-Mavrotheris and Lee attributed this difficulty to the fact that the raw data is transformed into a new form of representation. Even though this kind of interpretation of the result may be valid, it is possible to provide an alternate explanation for the unexpected results. For example, Wild and Pfannkuch (1999) emphasized the importance of providing students with meaningful ways to understand variation and seeing variation for a purpose and in understanding the consequences that variation can bring. Also, Reading and Shaughnessy (2004) specifically recommended that "different tasks or questions can encourage different forms of reasoning; for a chance to develop all aspects of their reasoning, students need to be offered the opportunity to react to a variety of tasks and respond to a variety of questions" (p. 223).

**10]** Suppose that students in five hypothetical statistics classes (Class F, Class G, Glass H, Class I, Class J) were asked to rate the value of statistics on a 1-9 scale. The ratings of each of the classes are shown in the following histograms.




| Rating | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Class ___ count | 0 | 3 | 1 | 5 | 7 | 2 | 4 | 2 | 0 |
| Class ___ count | 1 | 2 | 3 | 4 | 5 | 4 | 3 | 2 | 1 |
| Class ___ count | 1 | 0 | 0 | 0 | 22 | 0 | 0 | 0 | 1 |
| Class ___ count | 12 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 12 |
| Class ___ count | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Please fill out the table by putting down the name of the class corresponding to each of the five sets of counts in the table.

Q2. Judging from the tables and histograms, take a guess as to which has more variability between classes F and G.
Q3. Judging from the tables and histograms, which would you say has the most variability among class H, I, and J? Which would you say has the least variability?

Q4. Use Fathom and the data in HypoValue.ftm to check that you filled the table correctly.
Q5. Use Fathom and the data in HypoValue.ftm to calculate the range, interquartile range, and standard deviation of the ratings for each class. Record the results in the table:
Q6. Judging from these statistics, which measure spread, does class F or G have more variability? Was your expectation in (2) correct?
Q7. Judging from these statistics, which among classes H, I, and J have the most variability? Was your expectation in (3) correct?
Q8. Between classes F and G, which has more "bumpiness" or unevenness? Does the class have more or less variability than the others?
Q9. Among class H, I, and J, which distribution has the most distinct values? Does that class have the most variability of the three?
Q10. Based on the previous two questions, does either "bumpiness" or "variety" relate directly to the concept of variability? Explain.

*Figure 2.1.* "Value of Statistics" Task. Adapted from "Exploring introductory statistics students' understanding of variation in histograms" by M. Meletiou-Mavrotheris and C. Lee, 2005, in a paper presented at the Fourth Congress of ERME, the European Society for Research in Mathematics Education, Sant Feliu de Guíxols, Spain, p. 5.

A different approach to investigate students' understanding of variation in graphical representations is to ask students to draw a graph that shows a certain kind of variation (i.e., a little or a lot of variation). Based on the JLS, as discussed above, Garfield, delMas & Chance (2007) gave three lessons in an introductory statistics course at the University of Minnesota - one during spring semester and two during the fall semester. In the spring lesson, the authors asked students to discuss possible variables that had a little or a lot of variability in reference to graduating seniors at the University of Minnesota. Students then had to draw graphs for the variables they had chosen, and explain the reasoning for such graphs. After discussing with the entire class their graphs and chosen variables, students were guided to take a sample of actual data from graduating seniors of the University of Minnesota and using DataDesk software draw graphs and compare their predicted graphs with the graphs from real data. The results of this exercise indicated to the authors that students tended to focus on single data points (i.e. outliers) in a distribution.

Even though the lesson was intended to help students explore a measure of variability, Garfield et al. realized that the idea of distribution needed to be a strong foundation before students are able to appropriately deal with measures of variability. In addition to videotaping this lesson, researchers administered the Comprehensive Assessment of Outcomes in Statistics (CAOS) test (Garfield, delMas & Chance, 2002) before the course began and also at the end of the course (a slightly modified version of CAOS). They used the same test in all sections of the introductory statistics course and the test was also administered in two other

introductory statistics courses with the purpose of comparing results across different teachers and types of students.

The result of the tests at the end of course where the JLS process was used revealed that there was not much difference in students' understanding of histograms and variation from the pre-test and/or across sections and teachers. The main finding was that students were not able to connect the basic ideas of variability across topics as the researchers' intervention had hoped to establish.

After subsequent meetings, researchers designed two more lessons. Only the third lesson seemed to dramatically improve students' understanding of variation. The main difference in the third lesson was that Garfield et al. provided students with the summary statistics that accompanied the list of variables and their corresponding graphs. Students were then asked to rank the variables from the least to most variability and to justify their responses. Students were also asked to explain if data sets had different variability by looking at two of the data sets provided that had the same standard deviation, but different IQR (interquartile range). According to the authors, this third lesson improved students' understanding of variation as shown in the results of the post-test. For example, 94% of the students were able to correctly answer questions that asked them to compare two histograms that differed both in center and amount of variability. Another positive outcome of this lesson was that students were able to compare and describe variability of two data sets better at the end of the semester of class discussions.

Even though these results seemed more promising than the previous two lessons, Garfield et al. explained, that "it was very difficult to help the students develop the concept of variability as spread from the center" (p. 141). The authors recognized that even after several class sections on this topic, homework, and well thought out and designed activities, students still did not understand what the interquartile range and standard deviation represented or how they related to the idea of center and distribution.

A study conducted by delMas and Liu (2005) investigated the interplay that exists between graphical representation of a distribution and the numerical summaries, like the standard deviation. They believed that exposing students to activities that showed the relationship that exists between numerical summaries and graphical representation would produce not only a better understanding of variability in a distribution, but also provide a better foundation to the understanding of a measure of variability widely used like the standard deviation (SD). By being able to understand how the possible values of a variable and their respective frequencies with respect to the mean affect the SD, students could anticipate the value of the SD (p. 88).

delMas and Liu argued that numerical representations of variation like standard deviation (SD) are taught and required of students at all levels; however, the conceptual understanding of SD is difficult for students to fully comprehend (Shaughnessy, 1997; Reading & Shaughnessy, 2004).

In the delMas and Liu (2005) investigation, college students explored through a computer program how manipulating the bars of histograms affected the value of SD. Their study consisted of games that asked students to manipulate the distributions so that they would obtain the distributions with the largest SD. Another game consisted of finding different arrangements of the bars so that the distribution produced the same SD. Yet another game consisted of students manipulating histograms in such a way that they produced the smallest SD in each of three different distributions. The goal of these games was to assess students' understanding of how the changes in the distribution affect the SD.

The results of the delMas and Liu study revealed that students who were exposed to this intervention were able to recognize more aspects of SD than those who had not been exposed to the intervention. Specifically, students were able to recognize that mirror distributions would produce the same SD; that the relative and not the absolute location of the bars determined the SD; moreover, students realized that a large number of values needed to cluster around the mean to produce a relative small SD and that spreading the values away from the mean on both directions would produce the largest SD.

However, delMas and Liu were not too impressed with the results, as they stated that "only a few students provided justifications that partially resembled a fully coordinated conception of how frequency and deviation from the mean combine to influence the magnitude of the standard deviation" (p. 110). The results of this study left the researchers with the understanding that it seems very difficult to

fully comprehend SD. From this study, it can be concluded that even with a well-designed program and with the time allocated for an exploration of the graphs guided by highly qualified researchers, students fell short in understanding SD.

It is evident from the above-discussed studies that it is vital to fully grasp the concept of distribution as a prerequisite for understanding variation in graphical representations. It seems that giving opportunities to students to predict and draw graphical representation with different amounts of variability did not produce a greater understanding of variability. While in some studies it appears that students responded better when the relevant statistical summaries accompanied the graphical representations provided, in others the statistical summaries still proved insufficient in demonstrating more positive results in students' understanding of variation.

**Average – Connection to variation**

Since variability is a measure of the spread from the center, the conceptions that students (and instructors) have of the average can greatly affect the way they understand variation (delMas & Liu, 2007). However, it is known that average is mostly understood based on a formula for the average (Watson, 2007).

Watson (2007) explored the difficulties associated with the meaning students assign to the arithmetic mean. She found that only 30% of the students ($3^{rd}$, $6^{th}$ and $9^{th}$ graders, a total of 58 students) could solve more sophisticated applications of the average. For example, students would not give a reasonable explanation when they were asked to explain what it means that ten families had on average 2.3 children.

She credited this difficulty to the fact that most curricula teach only the average algorithm. Many students can state the algorithm, but they are unable to use it correctly and flexibly. She suggests that students need to be exposed to many representations and applications of the arithmetic average.

Recognizing the connection of the average with variability is not widely understood, since the average is mostly considered a single value to represent the data or the population. However, when collecting data, particularly large data sets, it becomes evident that most values of the data set will concentrate around a particular value and then there will be variation around that value. Konold and Pollatsek (2004) described the average as recognizing the "signal" or actual value of the center of the measurements and the "noise" as the variation around that value (p. 184).

Konold and Pollatsek (2004) described four different interpretations of the average: as a data reduction, as fair share, as a typical value, and as signal and noise. Data reduction, they explained, refers to boiling down a set of numbers into one value. This may suggest that some information is lost in that reduction. This indicates that as the data set gets larger the single value of the average becomes less representative of the whole group since more information gets lost. Fair share quantities are distributed evenly among the individuals. This approach, the researchers noticed, is used mostly in elementary schools without making reference to the fact that they are actually finding the average (p. 180). Typical value refers to the value occurring most frequently, which is a representation of the mode instead of the average (p. 178).

Finally, signal and noise is an interpretation of average in which each observation is an estimate of an unknown but specific value. Konold and Pollatsek (2004), to explain more clearly what they mean by signal and noise, gave the example of weighing a gold nugget 100 times on a pan balance to try to establish its weight when not all the measurements are exactly the same. These measurement differences can be attributed to some kind of error. Therefore, each measurement has a fixed component attributed to the weight of the nugget and a variable component which can be attributed to the imperfect measurement process. By carefully recording these values, they should all concentrate around the actual weight, which is considered the signal, and the errors are the variation around the signal, which they called the noise. In this way, it is easy to see the average "as a property of the object while viewing the variability as a property of a distinctly independent measurement process" (p. 184).

Konold and Pollatsek (2004) also gave the example of the early astronomers who wanted to calculate the position of a star, but their problem was that observed coordinates varied from observation to observation. Astronomers were hesitant to make use of an average observation and they also refused to combine their observations with other astronomers, as they thought that combining them would multiply instead of reduce the effect of the errors. After the mid-eighteenth century, astronomers began to use the average of multiple observations as they realized how those values stabilized on a particular value as the observations increased (p. 183). Therefore, understanding the average from only its algorithmic representation is a

limiting view of the concept that leads students to further difficulties with more advanced concepts, such as SD or comparing two groups.

**Variation and inference – the hidden background needed**

Inference is a central topic of statistics instruction because it enables students to review results of research studies that use inference and because drawing inferences from data is part of everyday life (Zieffler, Garfield, delMas & Reading, 2008). One of the main goals of inference is to go beyond the data at hand and draw conclusions about the population (Rossman, 2008). This central topic (inference) requires students' mastery of several statistics concepts such as distribution, sampling, repeated samples, and theoretical probability distributions.

As Saldanha & Thompson (2007) put it "drawing an inference from an individual sample to a population can not be understood deeply without reconciling the ideas of sample-to-sample variability and relative frequency patterns that emerge in collections of values of a sample statistic" (p. 275). Because statistical inference integrates many important ideas in statistics such as data representation, measures of center and variation, the normal distribution, and sampling (Zieffler et al., 2008, p. 46), it is important to understand how students understand these related topics.

Dealing with variability of the sample and making inferences about the entire population proves to have many challenges for students (Aquilonius, 2005; Nickerson, 2000; Thompson, Liu & Saldanha 2007). Even when students are able to perform the procedures of hypothesis testing correctly, they are be unable to use these procedures appropriately in applications related to their field of study (Gardner

& Hudson, 1999). It is recognized in the literature that "inference is so hard that even professional researchers use it inappropriately" (Erickson, 2006, p. 1).

Given that understanding inference depends greatly on the mastery of several statistics concepts, the following section of the literature review discusses results of studies dealing with understanding variability in repeated samples, in sampling distributions and in comparing distributions. The section ends with an overview of studies designed to investigate students' reasoning when dealing with informal inference analysis. As a result of the difficulties students face with the above-mentioned statistical concepts, researchers are moving away from formal inference and are exploring informal inference more deeply as an intermediary step to formal inference. This section of the review will therefore finish with a discussion of studies that investigated the benefits of informal inference.

***Variability in repeated samples.***

Shaughnessy, Ciancetta, & Canada (2004) investigated the written responses of 272 students (84 middle school, 188 secondary school, grades 6 - 12) when dealing with repeated samples. Their study consisted of providing middle school students with a mixture of 100 candies - 60 red and 40 yellow - and high school students with a mixture of 1,000 candies – 600 red and 400 yellow. Students were then asked to predict how many red candies they would get in a handful of 10 candies (100 for middle school students). The researchers then asked the question in a way that would motivate students to provide some variation in their responses. For

example, students were asked to predict what would happen each time if they repeated the process six times (the candies were replaced after each trial). Students' predictions were categorized as reasonable (3,7,6,6,4,6 or 7,5,6,8,3,6 or 5,6,6,7,6,6), too low (2,3,4,3,4,4), too high (6,7,8,8,7,9), too narrow (6,6,6,6,6,6), or too wide (0,1,4,7,9,10).

Most student responses to the first question provided only a single possible value instead of providing a possible range of values. This kind of response was credited to the fact that the question itself tended not to raise the role of variability in sampling (p. 183). The results to the second question about repeating the process six times revealed that 65% of the secondary students did not have a good feel for what would be likely to occur in six separate handfuls (p. 183).

A similar experiment was conducted by Reading and Shaughnessy (2004) who indicated that the concept of probability, as a value point answer, can jeopardize the concept of variability when doing repeated samples. "Probability questions just beg students to provide a point-value response and thus tend to mask the issue of the variation that can occur if experiments are repeated" (p. 208). In their experiment they surveyed 400 students grades 4-6 in Australia, and 700 students grades 9-12 in the United States. Their candy mixture had 100 candies - 50 red, 30 yellow and 20 blue. The result of this investigation indicated that less than 30% of all students surveyed were able to successfully integrate the roles of both center and spread in sampling scenarios. Also, one of their conjectures worth mentioning is that they believed that results like 5,5,5,5,5,5 were a consequence of students' familiarity with

probability. Reading and Shaughnessy (2004) stated that students are used to responding to probability questions like "what is the probability that…" or "what is the most likely outcome" (p. 210). As the probability of a red candy was 50%, the answer of five red candies for each of the six handfuls reflects this probability in each hand.

***Sampling distribution.***

In an attempt to discover senior high school students' emphasis in sampling distribution situations, Rubin, Bruce & Tenny (1990) presented 12 high school students with six questions. Their goal was to understand if students tended to focus on sample representativeness or sample variability. They described sample representativeness as the notion that a sample taken from a population will have characteristics similar to those in that population. They described sample variability as the notion that samples from a single population differ and thus do not match the population. They argued that a sample provides the investigator with some information about the population "not nothing, not everything, but something" (p. 314).

Rubin et al.'s (1990) study consisted of giving students a scenario in which they had to predict a possible sample of six Gummy Bears that were selected from a large container that had two million green and one million red Gummy Bears. Students were first asked to estimate how many green Gummy Bears might be in

their own package and then to estimate how many kids out of 100 would have the same number of Gummy Bears.

Their results indicated that students lack experience in thinking about distribution and they also found that students tended to not be consistent. In one scenario a student would focus on sample representativeness and on the other question students would focus on sample variability. Rubin et al. pointed out that students' challenge in understanding statistical inference comes from their inability to combine sample representativeness with sample variability into a single model of a distribution.

***Comparing distributions.***

The connection between distribution and variability is highly recognized in the literature (Cobb, 1999; Konold & Pollatsek, 2002; Reading & Reid 2004, Watson & Kelly, 2002). Bakker & Gravemeijer, **(**2004) stated, "without variation, there is no distribution, and without sampling there are mostly no data." (p. 149). Several studies have reported students' difficulties when making informal inference from a sample distribution about the population distribution (Chance, delMas & Garfield 2004; Leavy, 2006; Pfannkuch, 2006; Pfannkuch & Reading, 2006). Shaughnessy, Ciancetta, Best & Canada (2004), as part of a larger project, interviewed 24 students selected at random from a larger group of students. The number of students in the larger group is not stated in this study; however, they mentioned that ten classrooms participated from six different schools; two middle schools and four high schools.

Shaughnessy et al. (2004) gave students a pre-survey before instruction, then one interview after the first class intervention and then a second interview between the second and third class intervention. The results of the second interview are described here since it dealt with comparing data sets. The students' interview took place after a teaching investigation in which students had the opportunity to explore some sampling activities. Since the interview took place during the following semester, the researchers presented the interview students with Figure 2.2 as a reminder of what they had done previously in class with actual distributions.

When students were presented with Figure 2.3, 40% of the students could identify four of the graphs correctly. The rest of the students were given the opportunity to rethink their answers after the interviewer pointed out that two of the graphs were made up. With this knowledge, 54% of the students got four correct. Shaughnessy, et al. claimed that, even though sometimes the appropriate reasoning was not present with the correct responses, many students paid attention to appropriate aspects of the graphs when making decisions.

Overall, Shaughnessy, et al. explained that students had a tendency to rely mainly on extreme values. Also they stated, "students tend to believe that extreme values will occur a lot more often than they actually will" (p. 14). The researchers also pointed out that students did not focus on the center of the distributions because of the similarity of the center in the four graphs.

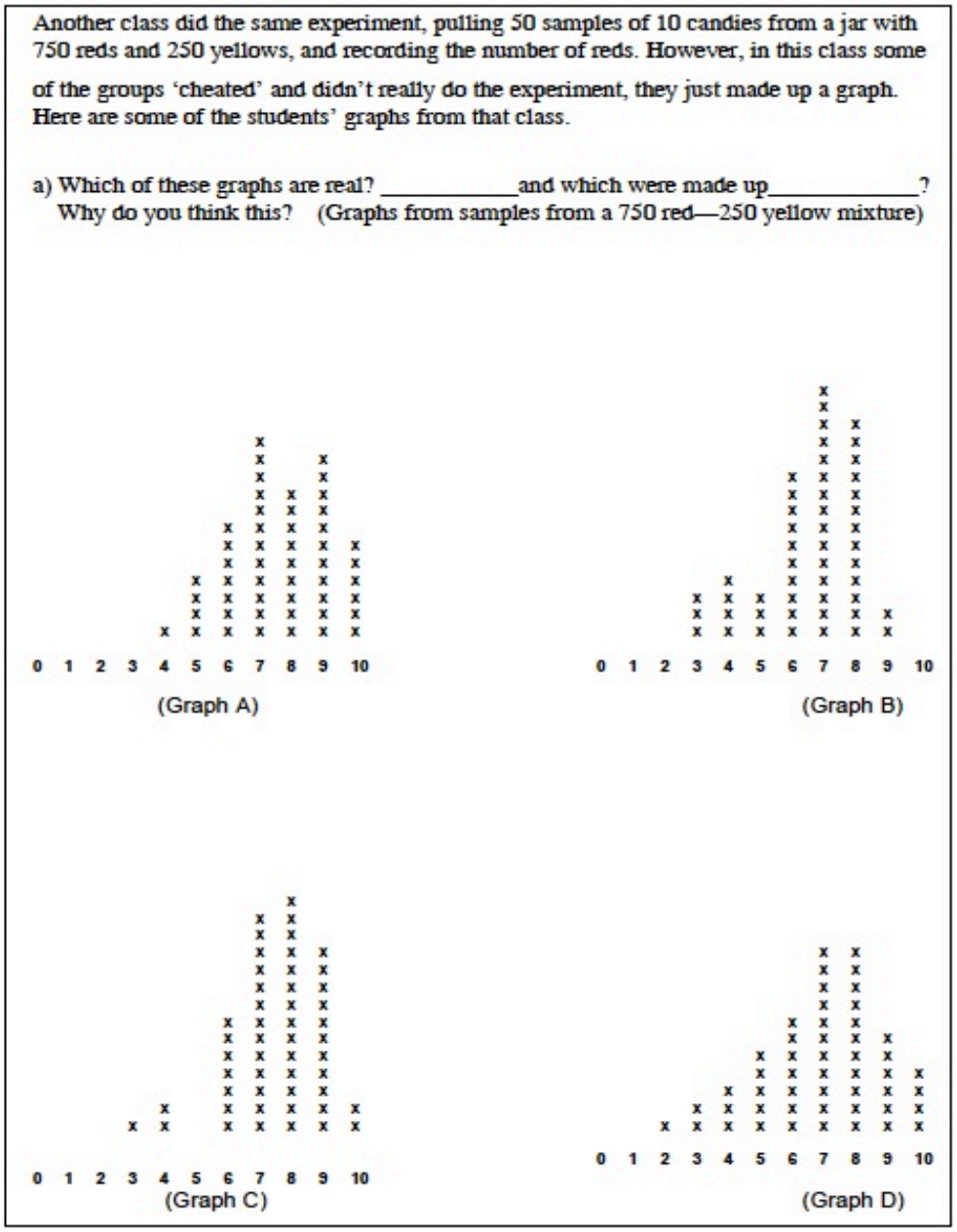



*Figure 2.2.* Four actual distributions. Adapted from "Students' attention to variability when comparing distributions," by M. J. Shaughnessy, M. Ciancetta, K. Best, and D. Canada, 2004, in paper presented at the 82nd Annual Meeting of the National Council of Teachers of Mathematics, Philadelphia, PA, p. 6.

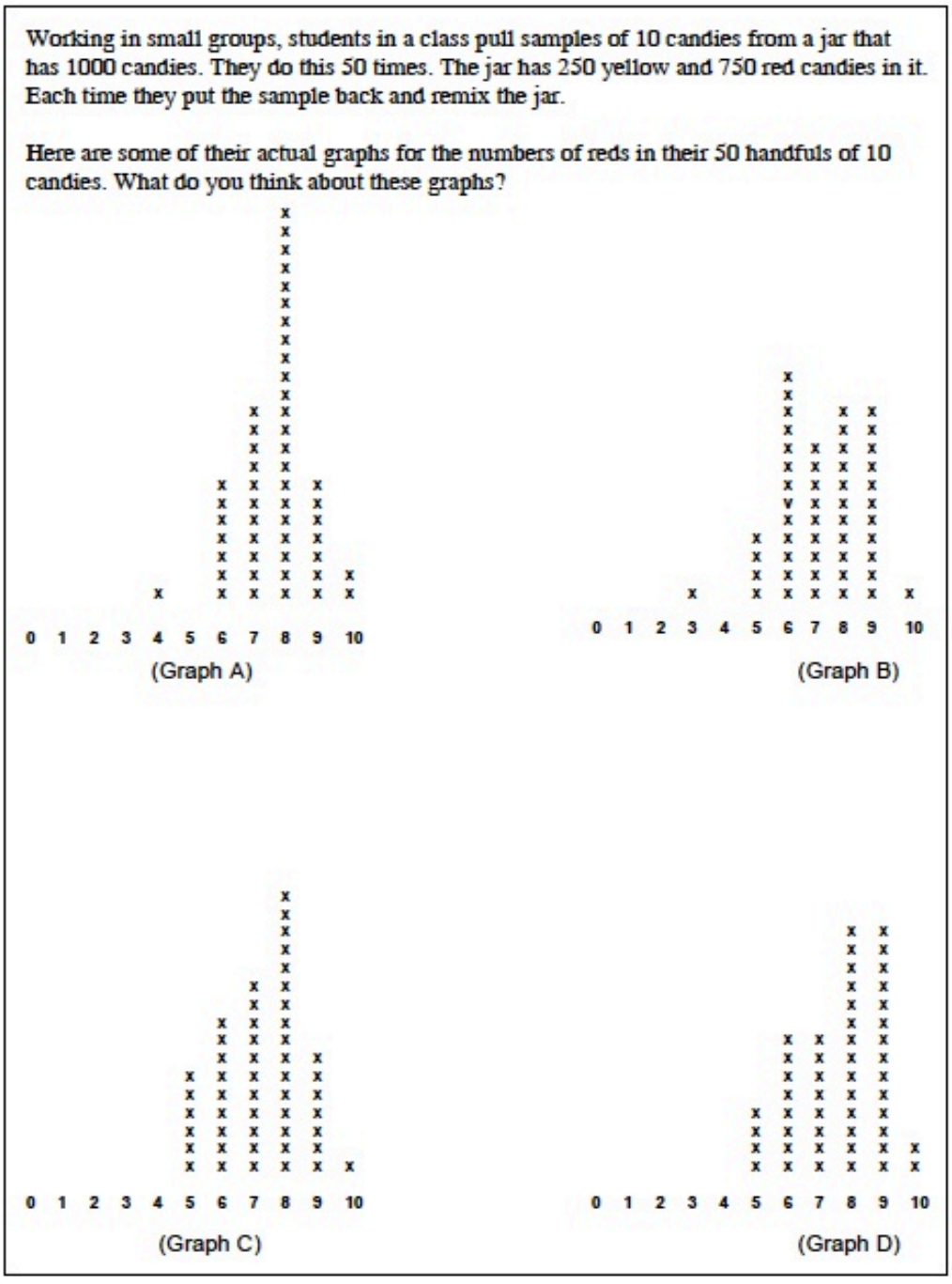



Figure 2.3. Which distributions are fake? Note that Graph B and Graph C are real and Graph A and Graph D are made up. Adapted from "Students' attention to variability when comparing distributions," by M. J. Shaughnessy, M. Ciancetta, K. Best, and D. Canada, 2004, in paper presented at the 82nd Annual Meeting of the National Council of Teachers of Mathematics, Philadelphia, PA, p. 7.

***Informal inference.***

"Inferential analysis is typically taught as a tool for judging the source of variation in data" (Pratt, Johnston-Wilder, Ainley, & Mason, 2008, p. 107). Traditionally, research on inference was dedicated to identifying students' struggles with concepts such as p-value (Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007) or to considering the complications students faced while deciding the appropriate test statistics when performing hypothesis testing (Quilici & Mayer, 2002). Research on formal inference has revealed that students struggle with the logic of statistical inference (Thompson, Liu, & Saldanha 2007). For instance, Aquilonius' (2005) analysis of college students identified that students did not fully grasp the relationship between the sample and the population in the context of hypothesis testing. She also found that college students had only a procedural view of p-value and therefore could not easily express the solution in the context of the problem. The author attributed this difficulty to the lack of understanding of the probability theory that lies behind hypothesis testing. She also noticed that textbooks and teachers provided students with scripted forms to answer, but this did not necessarily lead the student to the right solutions.

In contrast, statisticians, as per Wild and Pfannkuch (1999), are very involved from the beginning of the inquiry process and ask many questions, and also analyze graphs in many forms to gain as much information as possible before they decide what to do. In contrast, students are normally provided with a very narrow scenario and then asked to choose the best possible method. Biehler (1997) stated

that students need to develop a connected understanding and therefore “students have to overcome the belief that using one method or graph" is enough” (p. 174). Asking students to select one method or an appropriate statistics test with little information about the data forces the opposite of these recommendations.

Konold and Pollatsek (2004), in agreement with previous scholars, recommended moving away from the traditional statistics course which, they say, focuses on formal inference heavily based on fitting the data into mathematical models. They recommended a more exploratory emphasis where informal analysis is done before deciding which formal mathematical model is appropriate (p. 195). Slowly, researchers are stepping away from the traditional approach to inference and proposing informal inference as a path to formal inference. Makar and Rubin (2007) provided a clear distinction between formal and informal inference to understand the differences between them:

> Formal statistical inference, we refer to inference statements used to make point or interval estimates of population parameters, or formally test hypotheses (generalizations), using a method that is accepted by the statistics and research community. Informal statistical inference is a reasoned but informal process of creating or testing generalizations from data, that is, not necessarily through standard statistical procedures (p. 85).

A more formal definition has been offered by Zieffler et al. (2008) who state that “we present a working definition of informal inferential reasoning as the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples” (p. 44).

As a reflection of the new research approach to inference, the following section of this literature review will present studies that provide insight into what is meant by informal inference and provide some insight into its impact on student reasoning and understanding of statistics.

The study conducted by Pratt, Johnston-Wilder, Ainley, & Mason (2008) provided a good description of how inference requires the connection between sampling, repeated samples, distributions and graphical representations and comparing distributions. In an activity called "Guess My Die" students used a computer-based gadget to generate distributions and graphical representations. The authors wanted to assess students' difficulties with informal inference that connects data to probability distributions. Their goal was for the learners to restructure their attention from the local to the global. The researchers were interested in students' ability to draw inferences about the probability distribution associated with outcomes from a die. The study was conducted with groups of students of ages 10 to 11. The selection criterion as well as the number of groups who participated is not discussed in detail in this study, but the authors made reference to three different groups of students, two groups of two boys and one group of three girls.

In Pratt et al.'s (2008) study, students had to first guess the number of faces on a die when they were not aware of the underlying probability distribution. They were asked to focus on the sample data to deduce the population. For example, they had to decide whether the die was a fair six-sided die or if it was unfair. Or in some cases they had to decide how many faces the die had and the number on those faces.

Using the software called DataDesk created by Pratt, a pair of students was asked to guess "what is the die like?" The first die was not fair, but the students were not aware of it. So they drew samples of 32 trials and made pie graphs so that they could make an inference about the die.

Another of the activities utilized by Pratt et al. (2008) gave the opportunity for students to identify what the funny die looked like, i.e. how many faces the die had and what were the number on those faces based on the sample outcomes. Some dice had only five sides while others had 12 sides. The number of sides was only revealed to the students in the case of the 12 sided die, but not the five sided die. With the five sided die, students drew as many as 280 trials with the hope that they would get some stability or some confirmation for their deductions of the numbers on the die.

However, they seemed to be confused by the fact that no pie chart looked like the previous one and had great difficulty reaching their conclusions. The researchers had hoped that students would move from the local to the global approach and that they would recognize at least informally the benefits of the Law of Large Numbers. However, making reference to the 280 trials, one of the students said "just getting to stupid"; this means that he was less clear of the underling distribution as the number of trials increased. Therefore, his desire was to reduce the number of trial to 140 as he felt that less data would be better for his confidence in his conclusion.

Even though the results did not show the expected student outcomes, Pratt et al. are positive that this kind of gadget and interaction with the data can provide a

great link to more formal approach - “We conjecture that giving students the experience of mending gadgets before asking them to infer the nature of the die may be a necessary experience to enable students to understand more deeply the connection from data to modeling distribution” (p. 126).

The Pratt et al. (2008) study is relevant for investigating informal inference because students had no way of knowing the underlying population and therefore their expectations of what should happen were less influential. For example, with the fair six sided die, students had the preconception that a fair die needed to have the same number of outcomes for each of the side of the die, and they were confused with the variation encountered in a small number of trials. They could not ascertain the die was fair because of the variation. Students expected less variation in the outcomes to gain confidence in their results. One important aspect to consider is that these students were very young (10 to 11 years old). This study could easily be conducted with college students and instructors and the results may provide what authors were expecting in moving from the local to the global.

Another approach to investigating informal inference was introduced by a series of studies by Masnick, Klahr and Morris (2007). Their research focused on the way students’ reasoned about data and how they reasoned about variability to arrive at conclusions based on informal inference analysis. The results of several studies were combined to provide a uniting focus, namely how students discern whether the variability in the data is due to error or due to some effect that warrants a conclusion. These researchers also wanted to test whether prior held beliefs or knowledge

affected students' conclusions or if the characteristics of the data were influencing students' prior beliefs or knowledge.

The Masnick, Klahr & Morris (2007) study is unique since several ideas were explored simultaneously. First, they compared children's results (K-12) with adults' results (college students). Second, they used context in three different modes: very familiar context where prior beliefs are correct; not so familiar context where the prior beliefs were mostly incorrect; and third, no context so that conclusions would be based solely on data. Their main goal was to engage students in an experimental design to consider the effect that students' prior conceptions had on their understanding of variation and how they could discern variation that was evidence from variation that was simple noise. The authors wanted to compare the responses of students in three different scenarios: 1) students were asked to design an experiment with all its relevant parts (design, execute, measure, graph and interpret results) for a very familiar context; 2) the same approach was given to another group of students in which the context was not too familiar to students, but it was known to the researchers that some preconceived ideas students had about the context might influence students' thinking; 3) provide a different set of students with data with very little background information or context.

The Masnick et al. (2008) study involved about 220 students in total; 29 second graders and 20 fourth graders for a ball experiment; 49 children and 28 undergraduates for a pendulum experiment, and 44 sixth graders and 50 college students for the data without an experimental context.

The results of Masnick et al.'s study revealed that students were able to identify several sources of variation due to errors in measurement or execution, giving less recognition to the effects of errors caused by their faulty design. They concluded that when students' experiments matched their prior beliefs they were able to separate the signal (center) from the noise (spread) of the data. Because their prior belief was strong they disregarded the noise (spread), and were able to easily identify the center of the data (signal).

On the other hand, when the results of their experiment did not match their prior held beliefs, students were less certain of their conclusions and they could not easily distinguish between the variation due to error or due to real effect and identify the center of the data. The researchers also concluded that children were less able to use data to revise their prior theories than adults, particularly when the data did not agree with their prior expectations. Adults clearly differentiated between small variation due to random effect and variation due to real effect. For children, small random variation was confused with variation due to real effect because they could use it to confirm their prior beliefs. Therefore, Masnick et al. (2008) concluded that data affects theory and prior theories affect data interpretation

Masnick et al.'s results also revealed that students' confidence in their responses depended greatly on the amount of variability in the data. When comparing data sets students tended to provide decisive conclusions mostly when the data sets did not overlap. In summary, Masnick et al. (2008) concluded that children tend to recognize data variation in many experimental contexts. Even though

students had a difficult time articulating their justifications, they were able to use key features of the data in drawing conclusions.

The above-mentioned studies provide a glimpse of what is meant by informal inference. None of these studies required students to perform any formal computation or decide which particular statistics test to use. Students were very engaged with the data and therefore were able to use the logic of inference more naturally. Rossman (2008) explained that the logic of inference is similar to the modus tollens argument in logic with a probabilistic aspect added for good measure. He further explained his thinking with an example of two dice that produce mostly the sum of seven and 11. For example, if the dice were fair, then it would be very surprising to find a long list of seven and elevens. Since the observations produced long lists of sevens and elevens, there was strong evidence that the dice were not fair. Rossman (2008) argued that students can follow this modus tollens logic rather easily when dealing with a context that is familiar to them.

As discussed by Thompson et al. (2008), teachers had as many difficulties with inference as students, therefore, the approaches utilized by the above authors could be a starting point for developing activities to perform with those teaching statistics so that they too could become more aware of the logic behind inference and build up from informal inference to formal inference.

In this literature review, research results with instructors at the college level (Dabos, 2009; Garfield et al., 2007), as well as the high school level (Thompson et al., 2008), have demonstrated the many challenges these instructors face with

statistics concepts in general and with the concept of variation in particular. Moreover, studies on K-12 students (i.e. Shaughnessy, et al., 2004) as well as college students (i.e. Meletiou-Mavrotheris and Lee, 2005) have provided an overview of the complications associated with students' understanding of variation. The main message that emerged from these studies indicate is that in order to fully master the concept of variation students need a well-developed concept of distribution, which leads to connected topics like graphical representations, repeated samples, and sampling distributions. However, every one of these areas shows that much needs to be done in order to provide students with full mastery and comprehension.

Moreover, this literature review reveals new trends in statistics education, namely moving away from formal inference procedures into more exploratory data analysis as a way to develop informal inference. This area of research is still in its infancy, but the results show students getting a better grasp of the logic of inference, even if it is informal, and therefore may prove beneficial in the long run when dealing with formal inference. It therefore seems that exploring instructors' concepts on informal inference is step towards future improvements in teaching statistics. If those teaching statistics were made a priority in research, the results of such discoveries could trickle down to benefit students as well as produce exponential benefits to the field of statistics education. This focus on those teaching statistics seems like an excellent starting point for further research.

# Chapter III

## Methodology

This chapter describes the rationale for the procedures of data collection and data analysis. The first section delineates the theoretical framework that guided this investigation. This is followed by a description of the research design, including the selection of participants, location and instruments used in this study. It continues with an account of the steps followed for the data analysis and trustworthiness. This is followed by an explanation of how the research questions were addressed by these methods.

### Methodological Framework

The methodological framework that guided this study is based on a mixed research approach that involves the mixing of quantitative and qualitative methods. In this section, some of the characteristics of each of these two methods will be described, followed by an explanation of why the mixed model approach was chosen for this study.

According to Cresswell (1994),

> A qualitative study is defined as an inquiry process of understanding a social or human problem, based on building a complex, holistic picture, formed with words, reporting detailed views of informants and conducted in natural setting. Alternatively, a quantitative study is an inquiry into a social human problem, based on testing a theory composed of variables, measured

> with numbers, and analyzed with statistical procedures, in order to determine whether the predictive generalizations of the theory hold true (p. 232).

In other words, qualitative research places emphasis on looking closely at people's words, actions and records and examines patterns that emerge from the data. In contrast, a quantitative approach to research quantifies the results of the observations and overlooks the words.

Quantitative research can be used to describe what is observed in the data, or used to draw inference from the sample to the population. In this study it is used to compare the performance of the instructors based upon educational preparation and teaching experience. The main goal of this study is exploratory in nature, with the goal of exploring or researching instructors' conceptions of variation through instructors' written and verbal explanations. It was determined that describing two-year college instructors' current conceptions of variation is a way to lay the basic groundwork for further research, especially considering the paucity of this type of information in statistics education research.

The paradigm used in this study mixes both qualitative and quantitative research. There are two major types of mixed designed paradigms, namely mixed method research and mixed model research. In the mixed method research, the researcher uses the qualitative research paradigm for one phase of a research study, and the quantitative research paradigm for another phase of the study. However, in the mixed model research method, the researcher mixes both qualitative and quantitative research approaches across all stages of the research process. The mixed

model research paradigm was chosen because of the dynamic interactions that using both quantitative and qualitative tools bring into all aspects of the study. A numerical answer may show one level of understanding, but the words used in the justifications given by the participants reveal a much deeper level of understanding. The instruments used in this study to investigate instructors' conceptions of variation were designed with this in mind. The instruments had a questionnaire that was composed of open-ended or qualitative type items, as well as some close-ended or quantitative type of questions.

The benefits of utilizing mixed design are manifold. When two approaches are used to focus on the same phenomenon and they provide the same result, the resulting evidence is considered superior. Also, one method might highlight aspects that would have been missed if only a quantitative or a qualitative approach had been used. In this way, the mixed method paradigm has the capability of both corroborating and expanding the set of results.

## Research Design

### Participants.

The population of interest is two-year college mathematics instructors in California. Statistics courses are almost always taught in mathematics departments at this level and they are one of the most common courses offered in mathematics departments. Because of the high demand for these courses, any mathematics instructor might be called upon to teach statistics at some point in a career. A

purposeful sampling method was utilized because the sample is, as stated by Patton (2002), "information rich and illuminative, that is, it offers useful manifestations of the phenomenon of interest: sampling, then, is aimed at insight about the phenomenon, not empirical generalization from a sample to a population" (p. 40-41). This is in accordance with the main purpose of the study, which is to describe instructors' conceptions of variation and not to generalize. The purposeful sample chosen for this study was comprised of mathematics instructors who taught at two-year colleges in California.

Participants were recruited in several different ways. Six instructors at a local two-year college took part in a pilot study. Because the exact same methods were used for the final study, these six were included in the larger study. Multiple strategies were used to recruit the rest of the sample. The researcher attended three of the California Mathematics Council Community College conferences (CMC3). One was held in Monterrey, California, one in San Diego, California, and one in Riverside, California. Thirteen more instructors were added to the pool of participants from these conferences. The other strategy to get more volunteers to take part in this study was executed through an email sent to the mathematics department chairs of selected community colleges, asking them to forward a letter to their mathematics instructors that asked them to take part in the study. These colleges were chosen by drawing a 100 mile radius from the researcher's location and identifying all of the colleges in that area. Twenty-nine colleges were within the 100 miles radius.

At the end of the data collection process, which took more than a year to complete, a total of 52 professors from 33 different colleges were recruited to participate in the study. Their degree status and teaching experience are summarized in Table 3.1. All participants had a degree in mathematics because this is a California state requirement. As shown in the table, some also had a degree in statistics and they varied in terms of whether or not they had statistics teaching experience.

Table 3.1

*Number of Instructors Participating in the Study*

| | | Statistics Teaching Experience | |
|---|---|---|---|
| | | Yes | No |
| Mathematics Degree | Statistics Degree | 7 | 0 |
| | No Statistics Degree | 22 | 23 |
| Total Number of Participants | | 29 | 23 |

For the purposes of the analysis, the participants were divided into three distinct groups. The M group was the group of instructors who had only a degree in mathematics and had never taught statistics. The MT group was comprised of instructors who had a degree in mathematics and taught statistics often (M for mathematics and T for teaching). The ST group was designated for those who had a degree in mathematics, but also had a degree in statistics and taught statistics often.

Table 3.2

*Number of Instructors Who Took Each of The Instruments*

| | M | MT | ST | Total |
|---|---|---|---|---|
| Survey Only | 18 | 18 | 4 | 40 |
| Survey and Interview | 5 | 4 | 3 | 12 |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

Table 3.2 below shows the breakdown of the 52 instructors according to their background degree, their statistics teaching experience and which part of the study they participated in. About 56% of the instructors taught statistics and 44% did not.

**Instruments.**

Instructors' demographics were collected by asking a brief survey on the instructors' background. The survey included the following questions: How many years have you been teaching mathematics? Specify your background degree (mathematics, statistics, both or other). How often do you teach statistics (often, sometimes, never)? Two different instruments were utilized in this study to examine instructors' knowledge of variation: Survey (Instrument A), which was a survey that

had two formats: paper and online, and Interview (Instrument B), which was conducted by a talk aloud interview and had only one format, paper.

Both Instrument A and Instrument B represent a compilation of several items used extensively in previous empirical studies by known scholars, some of which were discussed in detail in the literature review. The items chosen were utilized in the past to investigate mostly diverse student populations, and some K-12 teachers' conceptions of variation. Some questions had been utilized extensively for many years to measure a particular concept, while others were used only once or twice in previous research: A.Q1 (Reading & Shaughnessy 2000, 2004; Shaughnessy, Ciancetta & Canada, 2004); A.Q2 (Canada, 2004); A.Q3 (Canada, 2004; Noll, 2007; Watson, Kelly, Callingham & Shaughnessy, 2003); A.Q4 (Canada, 2004); A.Q5 (Canada, 2004); A.Q6 (Reading & Reid, 2007); A.Q7 (Reading & Reid, 2007); A.Q9 (Watson et al., 2003; Sorto, 2004); A.Q10 (Meletiou-Mavrotheris & Lee, 2005); B.Q1 (Noll, 2007; Shaughnessy, Ciancetta, Best & Canada ,2004); B.Q2 (Reading & Reid, 2007); BQ3 (Comprehensive Assessment of Outcomes for a First Course in Statistics, CAOS, 2005); B.Q4 (Sorto, 2004); B.Q5 (Watson et al., 2003); BQ6 (Canada, 2004). None of these researchers, however, utilized these questions to investigate two-year college instructors.

The decision to include certain questions in the instrument to answer a research question was based on how the item appeared to bring out different conceptions of variation in previous research. The items on the instruments were designed to answer the research questions by revealing instructors' conceptions of

variation in several areas: probability, repeated samples, and sampling distribution, as well as their understanding of graphical representations and basic statistics. Table 3.4 shows the alignment of the questions asked in the instruments and the research questions. A question in Instrument A is labeled A.Q# and a question in Instrument B is labeled B.Q# and the research questions are labeled RQ#. Note that due to the omission of a key question in item A.Q8, this item has been excluded from Table 3.3 and was not included in the analysis done for this project.

Table 3.3

*Research Questions With The Corresponding Items From The Survey And Interview*

| | RQ1 | RQ2 | RQ3 |
|---|---|---|---|
| | Probability, repeated samples, and sampling distributions | Graphical representations and average of one data set | Graphical representations with different data sets |
| Survey Items | A.Q1<br>A.Q2<br>A.Q3 | A.Q5<br>A.Q9 | A.Q4<br>A.Q6<br>A.Q7<br>A.Q10 |
| Interview Items | B.Q1<br>B.Q2 | B.Q3<br>B.Q4<br>B.Q5 | |

*Note.* RQ1 = research question one, RQ2 = research question two, RQ3 = research question three.

Instrument A (Appendix A) had ten questions and Instrument B (Appendix B) had six questions. Besides the length, the main difference in the two instruments was to provide two different settings for instructors to express their conceptions as described below.

**Data collection procedures.**

Demographic information was collected in one of two ways. The first method was to verbally question the participants when the researcher first met them. The second method was to add the questions to the online survey as a prelude to Instrument A.

Two instruments (Survey and Interview) were used to investigate professors' conceptions of variation. Answering each instrument took approximately half an hour. The written survey (Instrument A) provided instructors with the freedom to think and write freely, and the talk aloud interview (Instrument B) resembled a teaching situation, i.e., students tend to ask questions which instructors have to answer on the spot, explaining their thinking to students. Survey (Instrument A) was administered to all instructors participating in the study, while Interview (Instrument B) was only given to the 12 participants who volunteered to take part in the talk aloud interview.

Instrument A was administered in two different ways: in person with a paper format with the researcher, and online. The online version was available for completion and submission through a website provided by the Social Science Survey Center at UCSB. A link was provided for instructors to answer the online survey. Both the online and the paper version of Instrument A were identically worded. The main difference was that those answering online could not go back to revise or fix any decisions or errors. This affected only one instructor who made a mistake on the

keyboard and she/he could not go back so she/he ended up with a few blank sections because of that mistake. This did not happen with the paper survey. When doing Survey (Instrument A), participants answered the ten questions without assistance from the researcher and they were asked to write down the justification for their choices.

Interview (Instrument B) was done in person with the researcher. Instrument B had exactly the same format (paper) for every participant and was answered in person with the researcher, and was both recorded and videotaped. It was always done after the instructors had completed Survey (Instrument A). Some instructors did it immediately following the completion of Instrument A. Others did it at a later date due to participants' time constraints.

With Instrument B, participants were asked to talk aloud as they responded to the six questions of this instrument. The idea behind this kind of format is that when instructors talk aloud and explain their thinking while solving a particular problem, they can more clearly express what they are thinking and therefore the researcher can discover interpretations that might be missed with a written format like that used for Instrument A (Van Someren, Barnard, & Sandberg, 1994). The researcher did not ask any questions of the participants. During data collection, the researcher did not interfere with instructors' responses, but she was available in the event questions arose. If participants asked questions, the protocol observed by the researcher was as follows: If the participant asked for validation of their responses, the researcher said "I can't answer any questions. State what things come to your mind to solve this kind

of problem", or the researcher said, "Explain what you are thinking." The researcher was consistent with all participants who experienced similar difficulties.

**Researcher's background.**

The researcher has a master's degree in statistics as well as a master's degree in education and a bachelor's degree in mathematics as well as a bachelor's in education. She has taught statistics and mathematics at a local community college for the last eleven years.

## Data Analysis

The data were organized by research question in alignment with the items designed to answer each question as shown in Table 3.3. Within a particular statistics topic, each question was first analyzed individually to allow the researcher to find patterns in the responses across the participants. Then the group of questions within the particular context was analyzed as it revealed an overall picture of the conceptions of variation held by instructors in that particular context.

Since the questions pertaining to Instrument A were answered by all instructors (n=52), they were analyzed in two ways. One was based on the numerical responses and the other by analyzing the words used to justify their thinking. Descriptive statistics was used to analyze the numerical responses. Pattern recognition was utilized to identify the different ways instructors justified their thinking. The coding of the instructors' reasoning will be discussed in the next

section. This method of organizing the data permitted the researcher to observe not only what the numerical values suggest, but also to uncover instructors' thinking. Hierarchies were not used in the coding of their explanations, but rather coding was used to identify the different ways to justify a particular response. Direct quotes from instructors are utilized in the results section of this study as evidence of their thinking. The last part of the data analysis was assigned to the coding of the responses to questions from Instrument B. There were only 12 instructors that answered Instrument B. This part aims to expand the vision of instructors' thinking, since instructors were able to express their thinking verbally instead of in writing like in Instrument A.

**Data transformation and representation.**

Once the data were collected, all the responses from Survey (Instrument A) - either the paper or online format - were transferred into an Excel file. The data from Instrument B were transcribed from the videotapes and the tapes were used as a second reference in the event that the voice on the video was not clear. All of the videos were watched again once the transcription was done, with the transcription in hand, to corroborate the transcription and for detailed information that was missed in the first transcription. The videotapes tended to be very useful in revealing visual clues. Moments of prolonged silence were transcribed with "..." or stating "silence"; also, if their verbal expression indicated thinking or insecurities these were also recorded; particularly when an instructor said "umm..." or "ah....". Moreover,

description of hand movement, i.e. pointing to a graph or section of a graph or drawing something was also noted in the transcript.

Once the data were compiled into one file, the data coding began. In qualitative studies coding and developing a category system are a major part of data analysis. Coding is defined as marking the segments of data with symbols, descriptive words, or category names. The type of coding used is called inductive coding, which employs codes that are developed by the researcher by directly examining the data (Johnson, 2006). The researcher carefully read the transcribed data, line by line, and divided the data into meaningful analytical units (i.e., segmenting the data). When she located meaningful segments, the researcher coded them consistently throughout all the participants' responses. The codes were reapplied to new segments of data each time an appropriate segment was encountered. The researcher continued with this process from participant to participant and from question to question until the researcher had segmented all of the data and had completed the initial coding.

During this process, the codes were emerging from the data and therefore some co-occurring codes were also developed. This means that the same lines of segments of text may have more than one code attached to them. The second coding was a subcategory of the first. The researcher kept a master list. After the researcher finished the initial coding of the data, she summarized and organized the data as she continued refining and revising the codes. The main goal of this coding was to identify similar kinds of responses by all of the participants. This led to exposing the

many different ways a question was answered. The coding did not establish hierarchies of the response but rather labeled an emerging pattern of the responses.

It is important to highlight that while prior research on probability and sampling has developed hierarchies of students' thinking based on the responses to the same or similar questions, the researcher decided not to employ such tools to measure instructors' thinking, but rather to expose the patterns of what they said. Trying to discern from a few written sentences what instructors really comprehend is to minimize their intellectual ability to a label, which seem highly inappropriate considering that the instructors who took part in this study have proven to be highly intellectually capable.

When analyzing the transcripts as well as the written answers of the survey, the researcher looked for patterns across the participants to summarize the findings into emerging categories. Some categories and results had previously emerged in the pilot study (Dabos, 2009) and were used as a lens for investigating the larger sample; even though new categories emerged this served as a starting point of investigation.

The analysis was followed by a comparison of the results according to instructors' education and statistics teaching background; for example, comparing responses across the M, MT, and ST categories. In some cases it was considered relevant to consider only instructors' statistics teaching experience by combining the responses of the MT and ST categories (all instructors who teach statistics) and comparing them to the M category (instructors who do not teach statistics).

The benefits of this type of consolidation and analysis included the ability to

highlight the differences between instructors' teaching statistics experience, and the ability to minimize small cell count. When instructors' justifications were separated into several categories, it was considered more appropriate to simply compare participant responses by statistics teaching experience, particularly given that the ST group consisted of only seven instructors. Also, due to the small number of participants (n=12) who answered the interview (Instrument B), comparing responses according to instructors' statistics experience (M vs.MT/ST) was considered more relevant than separating them into three small groups (M, MT and ST).

In the following three chapters (chapters four, five and six) the results of the study will be presented. The last chapter (chapter seven) contains the discussion of the results as well as implications for findings and directions for future research.

# Chapter IV

## Results Pertaining to Research Question One

The analysis of the results presented here aims to answer Research Question One - What are instructors' predictions in repeated sample experiments? - along with its sub-questions:

- Is sample variability reflected in instructors' predictions?
- Are instructors predicting an appropriate amount of variability?
- What statistical concepts do instructors focus on when dealing with empirical sampling distributions?
- How are teaching experience and education related to predictions in repeated samples?

First, the researcher presents the results of analyzing all 52 participants who answered questions from Instrument A, which deals with repeated samples (A.Q1, A.Q2, and A.Q3). Then, focus will be given to the results of the analysis of the 12 instructors who answered questions from Instrument B, which also deals with repeated samples (B.Q1 and B.Q2). The chapter ends with a discussion pertaining to these results.

The main purpose of A.Q1, A.Q2 and A.Q3 (Appendix A) was to gain insight into instructors' predictions when prompted with scenarios that deal with taking several samples from a known population, with replacement. The first question in this section (A.Q1) had several sections (a, b, and c) (Figure 4.1). In each

section, the instructors were asked a question that required a short answer response, and then asked to explain why they gave that particular response. These sections were designed to investigate how responses may change according to the wording of the question and to probe the reasoning behind the responses (Reading & Shaughnessy, 2004).

Part 'a' asked a question that should be answered with a range of values, but in previous research (Reading & Shaughnessy, 2004) it has been shown to produce only a single value type of response. This previous study investigated only students' understanding in this type of question; it was therefore considered an interesting question to ask instructors to see if they would give a range of numbers, as the question was meant to produce. Part 'b' guided instructors into thinking about sample variability, and part 'c' asked respondents to directly predict the possible values when taking six samples from the same population with replacement.

A similar scenario was introduced in the second question (A.2), where instructors were asked to predict a range of values when taking six different samples, and then 30 different samples. The third question (A.3) dealing with repeated samples was presented with a different scenario than the candies (A.Q1, A.Q2 - mixed candies, A.Q3 - rolling a fair die), but with the same purpose, namely, to examine instructors' thinking about variability when taking repeated samples.

Figure 4.1 Instrument A, Question 1. Coded as A.Q1

1] Suppose there is a container with 100 pieces of candy in it 60 are red and 40 are yellow. The candies are all mixed up in the container.

You reach in and pull out a handful of 10 candies at random.

(a) How many red candies do you think you might get?
Why do you think this?

(b) Suppose you do this several times (each time returning the previous handful of 10 candies and remixing the container).

Do you think this many reds would come out every time?

Why do you think this?

(c) Suppose six students do this experiment (each time returning the previous handful of 10 candies and remixing the container).
Write down the number of reds that you think each student might get:

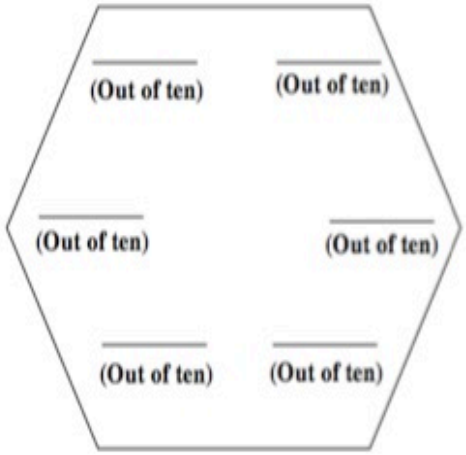



*Figure 4.1.* Instrument A, Question 1. Coded as A.Q1. Adapted from "Students perceptions of variation in a sampling situation" by C. Reading and J. M. Shaughnessy, 2000, in paper presented at the 24th Conference of the International Group for the Psychology of Mathematics Education, Hiroshima, Japan.

**Analysis of Question A.1**

Question A.1a was designed to provoke a range of values as response. However, the results show that for the most part, instructors did not provide a range of values but, as in previous studies, predicted a single numerical value. However, some instructors were able to explain in the written justification that they recognized the possibility of a range instead of a single value. Therefore, the analysis of instructors' reasoning provided another insight into instructors' thinking about variability when dealing with repeated samples.

Responses that included a single value - in this case the value six - were considered as not predicting variability in the sample and coded NPV. Responses that included a range of values (M11: 5, 6 or 7) were considered as predicting variability and coded PV. For this question (A.1a), 12% (n=6) of instructors predicted variability, giving responses like "4 to 8" (ST22), or "0 to 10" (MT49), while 88% did not predict variability. This reveals that almost all of the instructors focused on the center of the distribution or the expected value. This indicates that this question did not produce the desired outcome (a range of values) as it was intended. On the other hand, having instructors, however few, responding with a range of values brings to light that some people volunteered the idea of variability readily while others did not.

Instructors' justification for A.Q1a provided another layer of understanding instructors' reasoning when dealing with repeated samples. Their justifications were categorized into two main parts: Proportional Reasoning (PR) in which instructors

mentioned 'percent of reds', or 'probability of reds' in their justifications and thus making connections to population proportions, and Conception of Variation (CV) in which instructors exhibited explicit or implicit attention to either sample variability and/or spread from center.

Table 4.1

*Examples of Proportional Reasoning vs. Conception of Variation (A.Q1a)*

| ID | Response | Justification |
|---|---|---|
| Proportional Reasoning | | |
| M16 | 6 | I expect the proportion I get to be the same as the population proportion |
| MT40 | 6 | Because 60% of the container contains red candies. 60% of 10 = 6 |
| Conception of Variation | | |
| M11 | 5,6,7 | I would expect either 5,6 or 7 since empirical probability indicated 6/10 something close to but not necessarily equal to that seems likely |
| ST22 | 4 to 8 | 60% are red, so I'd expect a long term average of 6 red. However, due to randomness, I'd expect 6 "give or take" a few. |
| MT38 | 5 to 7 or 4 to 8 | Between 5-7… maybe 4-8 B/C theory & reality aren't the same |

*Note.* This table shows some instructors' written justifications that highlight both Implicit Proportional Reasoning found in 85%, (n=44) of participants, and Conception of Variation found in 8%, (n=4) of the participants.

Of all the instructors, only 8% demonstrated conceptions of variation, while 84% gave justifications that indicated proportional reasoning and 8% gave responses that were not interpretable. Examples of instructors' statements along with their numerical predictions are in Table 4.1. When looking at the justifications, it becomes

more evident that those who gave a range of values acknowledged sample variability, meaning that samples are likely to differ from the population. On the other hand, those who gave proportional reasoning focused only on the center of the distribution, ignoring other aspects of the distribution like spread, spread from the center and/or the shape.

By analyzing instructors' reasoning it was discovered that even though some instructors gave only a single value response in the numerical predictions, they still recognized the possibility of variation by expressing it in their written justifications; for example, MT18 said: "theoretically I should get 6 reds since 60% of the candies are red. But in actuality, I would probably get a different number". This indicates that this question brings up a tension between sample representativeness (sample mirrors population) and sample variability (sample differs from population).

Question A.1b guided instructors to think about what would happen in a repeated sample. The idea behind this question was that prompting participants to consider more than one sample should more readily trigger instructors to express ideas of variability in their responses. In fact, the majority (79%, n=41) recognized that the same number (six) would not appear every time if more samples were taken, indicating that instructors recognized sample variability. However, 21% (n=11) of instructors still decided that they would get the same value every time.

This is further demonstrated in responses given to question A.1c. This question probed instructors' understanding about repeated samples by asking them to predict the outcome for six repeated samples. The results show that even when

giving a correct response, instructors did not always use accurate explanations for these answers. The responses for the first part of this question (A.1c) were first coded for whether the answer demonstrated variability (different numbers for the six samples) or not (all numbers were the same). Then the responses showing variability were coded as either appropriate or inappropriate. Following Reading and Shaughnessy (2004), inappropriate responses were those that were too low (M25: 2,3,2,1,3,5), too high (M43: 5,6,7,8,9,10) or too broad (M52: 0 to 10). Appropriate responses were (M34: 5,5,6,7,7,8), or (M28: 3,4,5,5,6,7) or (MT: 5,6,6,7,7,7); basically, appropriate responses were those predictions that were neither too low, too high, nor too broad.

For this question (A.1c), 79% (n=41) of instructors gave responses indicating variability, while 21% (n=11) used just one number for the six samples, reflecting no indication of variability. When looking at those responses that indicated variability, 32% (n=13) of instructors gave answers that were defined as inappropriate, while the other 68% (n=28) gave answers that were considered reasonable. Overall, when the responses of all instructors are considered only 54% (n=28) of the instructors gave responses that were considered fully correct.

Instructors' written explanations provided additional information about their reasoning about repeated samples. For the most part, the justifications confirmed what the numerical responses showed, i.e. they either considered variability or not. However, some written explanations showed tension in instructors' thinking, while

others highlighted some of the instructors' limitations when considering repeated samples.

When looking at the statements accompanying the numerical predictions, it was found that, within the numerical responses that were labeled reasonable, 58% (n=18) of instructors provided justifications that showed awareness of the influence of spread from center, and 16% (n=5) expressed that variability was due to randomness. These types of reasoning reinforced what the numerical responses had shown, and provided greater depth in understanding instructors' conceptions of variation. However, other justifications showed more limitations in their conceptions of variation than the numerical predictions had suggested. Within those who gave reasonable numerical predictions, 26% (n=8) of instructors gave reasoning that emphasized only the expected value, or they stated that they had guessed, or that they did not know or they stated that anything is possible (see Table 4.2).

Table 4.2

*Examples of Reasoning that Shows Uncertainty and/or Focus Only on Range (A.Q1)*

| ID | Six Samples Predictions | Example of Instructor's Reasoning |
|---|---|---|
| MT49 | 456668 | Because the students could pick any number of reds between 0 and 10. |
| MT28 | 344679 | I really do not know! |
| M47 | 256678 | Truthfully I don't understand how I am supposed to be thinking about this question. I just picked numbers between 1 and 10. |

While 60% (n=31) of the instructors' numerical responses predicted reasonable variability, looking in greater depth into instructors' justifications for such predictions revealed that only 44% (n=23) of the 52 instructors demonstrated a sound conception of variation. It seems that the deeper the investigation, the less certain instructors' conceptions of variation appear to be.

Justifications for those who did not predict variability tended to confirm their focus on the expected values or the center of the distribution. Within those who did not predict variability, 80% (n=8) gave justifications that focused only on the center of the distribution, ignoring other aspects of the distribution. The other 20% (n=2) stated that they did not know or did not respond. The more interesting discovery in the justifications when juxtaposed with the numerical responses came from those who seemed to predict variability but whose justifications lacked depth.

Table 4.3

*Percent of Instructors Predicting Variability in Repeated Samples A.Q1*

| | | ST<br>n=7 | MT<br>n=22 | M<br>n=23 |
|---|---|---|---|---|
| A.Q1a | PV | 14% | 9% | 13% |
| | NPV | 86% | 91% | 87% |
| A.Q1b | PV | 86% | 72% | 82% |
| | NPV | 14% | 14% | 9% |
| | Other | 0% | 14% | 9% |
| A.Q1c | PV | 43% | 77% | 87% |
| | NPV | 57% | 18% | 13% |
| | Other | 0% | 5% | 0% |

*Note.* Results based on Short Answer Responses. PV indicates predicted variability and NPV did not predict variability. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

When the results were organized by instructors' statistics teaching experience and education, it was discovered that the percentages do not differ greatly for most of the question A.1 short answer responses (Table 4.3). Among the instructors who predicted variability in part 'c' there is a difference in the percentages between the groups. However, it is not the kind of difference that may intuitively be expected. The lowest percentage comes from those with a degree in statistics (note that the sample is small and therefore a statistical test cannot be used to infer about differences between the groups). The percentages shown in Table 4.3 reveal some interesting observations. For example, only one of the seven participants with a degree in statistics gave a range of values in response to A.Q1a. In addition, those who have never taught statistics are equally likely to predict variability as those who have. It is tempting to conclude that teaching experience and/or degree are not the determinants of the type of responses instructors provided in question A.1a; something else may be a factor, which needs further investigation.

**Summary of results of question A.1.**

As mentioned earlier, A.Q1 had several sections, namely a, b, and c and each of these sections had a short answer and a justification for such answer. From the short answers it was discovered that two-year college instructors' responses reflect, for the most part, what other studies have found (Reading and Shaughnessy, 2004) - it appears that the wording of the question altered the kind of response instructors gave. For example, this is highlighted in the contrast between 12% (n=6) of

instructors predicting variability in A.Q1a with 77% (n=40) predicting variability in A.Q1c. Additionally, in the short answer it was discovered that giving different numerical values for the six repeated samples did not necessarily indicate appropriate conception of variation, as some of the predicted values were considered too low, too high, or too broad.

Instructors' justifications revealed their conceptions of variation more in depth. While the majority seemed to understand that taking six samples would not provide the exact same values each time, there is some evidence, through the analysis of instructors' reasoning, that instructors' conceptions of sample variability appeared to have gaps. Words like "I don't know" reflect a lack of depth in their prediction of variability, and it is surprising to find statements like "Could be any number from 0 to 10. Other than that I have no idea" (ST2) from instructors who have a degree in statistics and who teach statistics often.

Their reasoning also highlights a division between those who readily predicted and justified sample variability and those who gave explanations that do not show that sample variability has been considered. However, the difference does not seem to be predictable by instructor degree or by statistics teaching experience.

Overall, the justifications behind the numerical responses provided by instructors reveal, at least at one level, their reasoning about variability. As expressed in their written statements, some instructors seem to have incorporated the appropriate conceptions of variability while others appear to show some limited points of view when it comes to taking samples from the population. The tension

between sample representativeness (same as population) and sample variability (different from population) seems evident in instructors' reasoning. Some stated that anything was possible (ignoring the distribution of the population), and others were grounded in the idea that the percentage had to mimic the population, ignoring sample variability.

The results of the combination of the short answers and their justifications demonstrate that it is important to pay attention not only to the numerical responses to a particular question, but also to instructors' reasoning behind those numbers, since they reveal aspects of their reasoning that the numerical values hide.

**Analysis of Question A.2**

The analysis of questions dealing with repeated samples continues with A.Q2, (Figure 4.2) which was designed so that instructors would not focus so much on a point estimate, but instead give a range of possible values, and in this way predict spread of the data and acknowledge some kind of variability. The results show that a majority of instructors recognized that taking more samples would make the overall range wider; however, their justifications demonstrated uncertainty as well as contradictory statements.

While coding the responses of this item, the researcher paid attention to whether instructors mentioned any or all of the three main characteristics that are manifested by increasing the number of samples taken from the same population, namely: 1) the overall range will be wider, 2) the overall cumulative values will be

closer to the expected value, and 3) it will better reflect the shape of the underlying distribution.

**Figure 4.2 Instrument A, Question Number 2**

**2]** Suppose 6 people did this experiment – pulled ten candies from the container, wrote down the number of reds, then returned the ten and remixed all the candies.

a) What do you think the number of reds will most likely to go from?

From a low of _____ to a high of _____.

b) Now suppose 30 people did this experiment. What do you think the number of reds will most likely go from?

From a low of _____ to a high of_____

c) Why do you think this?

*Figure 4. 2.* Instrument A Question 2. Adapted from "Elementary preservice teachers' conceptions of variation" by D. Canada, 2004, Portland State University, Portland.

The results from Question A.2a and A.2b indicate that 62% (n=32) of the instructors predicted that the range would be wider when the number of samples changed from six to 30 samples, and that a large percentage of instructors recognized that with more samples the possibility for extreme values increases, making the range wider. The rest of the instructors showed some difficulty in their estimations:

25% (n=13) of the instructors gave ranges that did not vary from six samples to 30 samples, 4% (n=2) provided ranges that were lower in the 30 samples than in the six samples, another 4% (n=2) gave values that were not interpretable, and 6% (n=3) did not respond to this question.

Within those who gave the same range of values (25%, n=13) for both six and 30 samples, there were two types of responses; one in which instructors (46%, n=6) predicted the range to be '0 to 10', and one where 54% (n=7) included a particular range close to the center but that did not vary from six to 30 samples (MT28: 4 to 8). It is easier to understand those who gave the sample space (0 to 10) as a range for both cases (six samples vs. 30 samples) since the sample space represents all the possible values of an outcome in an experiment. On the other hand, it is not clear how instructors came up with a particular fixed range like 4 to 8 (MT28).

The analysis of instructors' reasoning reveals that none of the instructors made reference to how taking more samples would reflect the shape of the population. This may be due to the question itself, which asked only for a range of values. The majority of instructors who gave a wider range of values for the 30 samples clearly identified that the possibility of extreme values increases when more samples are taken. Those who provided the same range were either confused or provided statements that seemed contradictory.

Table 4.4

*Examples of Predictions of Range Values for Six and Thirty Samples and Corresponding Justification A.Q2*

| | Six samples | 30 samples | Instructors' Reasoning | ID |
|---|---|---|---|---|
| Same | 0 to 10 | 0 to 10 | Could be any number from 0 to 10. I have no idea what the actual probability are without doing some actual calculation. | ST2 |
| | 0 to 10 | 0 to 10 | I'm not sure what you mean by "most likely." I'm interpreting it as "possible." | MT40 |
| | 4—8 | 4—8 | The proportion of red is .6, and the selection is random. | MT46 |
| Wider | 3—9 | 2—10 | With more people, there would be more chances for high or low results. The chances of some one pulling out 10 yellows is still quite low, and at a guess, I would say it would be fairly unlikely that even 1 person in 30 would see that result. 10 reds, however, seems more reasonable | M16 |
| | 4—8 | 2—9 | Outcomes which are farther away from 6 will occasionally, and rarely, show up. The more times we do this, the more likely it is that we will see an unusual result. | MT8 |
| | 4—8 | 3—9 | The more samples we take, the more extremes are possible. | ST48 |

Table 4.4 shows examples of instructors' reasoning according to the type of range they predicted. Instructors who predicted the same range to be '0 to 10' gave reasoning that transmitted uncertainty either about what the question was asking or with their predictions. Also, within those who gave the same range of values are

explanations that seem contradictory, for example, giving the same range (4 to 8) for both scenarios and stating that those values are "random" (MT46). It is not clear how the word random is used in this context.

Table 4.5

*Short Answer Responses to A.Q2. Type of Range Predictions by Instructors' Group*

| Range predictions | M | MT | ST |
|---|---|---|---|
| From six to 30 samples | n=23 | n=22 | n=7 |
| Wider | 65% | 55% | 71% |
| Same | 26% | 23% | 29% |
| Lower | 4% | 4% | 0% |
| Not interpretable | 5% | 4% | 0% |
| No response | 0% | 14% | 0% |

*Note.* Instructors' group are defined by education and teaching experience. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

When considering responses according to instructors' education and teaching experience (Table 4.5), it becomes evident there is no great difference in the type of responses given by instructors to the short answer section of Question 2. When paying attention to instructors' reasoning it is unexpected to find instructors who have a degree in statistics giving statements that reflect a lack of awareness for the empirical sampling distribution. For example, ST2 stated, "could be any number

from 0 to 10. I have no idea what the actual probability are without doing some actual calculation". However, this question did not ask for any probability.

**Analysis of Question A.3**

The last question that dealt with repeated samples was A.Q3, which asked instructors to predict the number of times each number on a die showed up in sixty samples. As with question A.1c, the results show that even when instructors gave a correct response, they did not always use accurate explanations for those answers. Similarly, in several cases where instructors gave incorrect responses for the short part of the question, their justifications expressed correct conceptions of variation. This indicates that the numerical answers may not necessarily reflect instructors' conceptions of variation.

Responses that included the same number for each face of the die (10,10,10,10,10,10) were coded as not predicting variability, and responses that included different numbers for each of the six faces of the dice (9,12,11,10,8,10) were coded as predicting variability.

For this question, 27% (n=14) of the instructors gave responses indicating variability, while 69% (n=36) used the same number for each face of the die, reflecting no consideration of variability in the samples, and 4% (n=2) did not give an answer to this question. Moreover, when looking at the responses to both questions (A.1c and A.3), it is found that only 25% (n=13) of instructors predicted variability in both, while 19% of instructors did not predict variability in either of these questions. Fifty percent predicted variability in the first question but not in the

second, and 6% did not respond. It is evident that there is a small percentage of instructors who consistently predicted variability across questions.

It becomes evident by looking at the justifications given by instructors that there is some tension in the way instructors think about this question. About 27% (n=14) of instructors' numerical responses and their justifications did not match. For example, 8% (n=4) gave predictions that showed variability when they predicted the numbers for the sixty samples but the justifications were not interpretable or were contradictory. For example, MT37 predicted (8,11,10,13,9,9) for the short answer and wrote "B/C '4' is my lucky number", and MT31 predicted (10,9,11,12,9,9) for the short answer and wrote "because it's a fair die, so every face has the same chance to show".

Some instructors gave predictions that did not show variability in the numbers, but gave justifications that acknowledged sample variability (19%, n=10). Table 4.6 shows examples of such justifications. Two things can be deduced from this type of tension: one, the numerical answers do not clearly reflect instructors' thinking; two, the tension between sample representativeness (sample reflects the population) and sample variability (sample differs from the population) continues to appear in instructors' responses. Some instructors were consistent in both parts of the question (Table 4.7); either by acknowledging variability in the numbers as well as the justifications, or by writing values in the numerical response that did not show variability and then justifying those numbers as representing the expected values.

Table 4.6

*Comparing Numerical Responses with the Justifications A.Q3*

| | ID | Numerical Predictions | Justifications |
|---|---|---|---|
| Tension | MT31 | 10,9,11,12,9,9 | Because it's a fair die, so every face has the same chance to show. |
| | MT49 | 10,10,10,10,10,10 | No, not really, the numbers could come up in any amount, but the expected number is 10 |
| Acknowledge variability in both | ST1 | 14,12,8,10, 7, 9 | Seems to be reasonable amount of sample variability |
| | M34 | 12,11,9,10,8,10 | It would be completely random with the probabilities approaching 1/6 for each number- but in a real sample there would be variation |
| No consideration for variability in either | ST3 | 10,10,10,10,10,10 | Fair die= uniformly distributed |
| | MT | 10,10,10,10,10,10 | Each number has an equal chance of being tossed. |

*Note.* 27% (n=14) of instructors showed tension, 19% (n=10) of instructors acknowledged variability in both, 50% (n=26) of instructors in neither, and 4% (n=2) of instructors did not respond.

When trying to compare short answer responses to A.Q3 by instructors' education and teaching statistics experience, it becomes evident (Table 4.7) that there is no great difference in the type of numerical response given by instructors. Finding that 65% of instructors who did not predict variability had no statistics teaching experience is not a surprise, as it may be justified by their lack of experience with the idea of sampling. However, discovering that 68% of those who teach statistics and 86% of those who have a degree in statistics did not predict variation in this item is somewhat more puzzling. It could be argued that the wording

of the question could be a determinant for the lack of variability prediction in this item. However, the 27% (n=14) of instructors who did predict variability read the same question as those who did not predict variability. So, some instructors seemed to volunteer responses that include sample variability regardless of the wording of the questions and others did not.

Table 4.7

*Percentage of Responses to the Numerical Part of A.Q3*

| | ST<br>n=7 | MT<br>n=22 | M<br>n=23 |
|---|---|---|---|
| PV | 14% | 32% | 26% |
| NPV | 86% | 68% | 65% |
| No response | 0% | 0% | 9% |

*Note.* PV indicates that instructors predicted variability by giving numbers that differed from the expected value. NPV indicates that instructors predicted variability by giving numbers that represented the expected value. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

Given that questions A.1c and A.3 asked for a similar type of response, it is interesting to look at the consistency of the numerical responses given to questions A.1c and A.3 by instructors' education and statistics teaching experience (Table 4.8). Again, it is evident that the percentages do not differ greatly.

Table 4.8

*Percent of Responses by Group of A.Q1c and A.Q3a*

| A.Q1c | A.Q3 | M n=23 | MT n=22 | ST n=7 | All n=52 |
|---|---|---|---|---|---|
| PV | PV | 26% | 27% | 14% | 25% |
| PV | NPV | 52% | 50% | 43% | 50% |
| NPV | NPV | 13% | 18% | 43% | 19% |

*Note.* Results based on Short Answer Responses. PV indicates instructors predicted variability; NPV indicates instructors did not predict variability. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics. Note the percentages do not add up to 100% because 3 instructors (2M and 1MT) did not respond to one of the two questions A.1c or A.3

Not only do the predicted short answer responses seem similar between the groups, but the reasoning behind them also bears similarity. For example, ten participants explained that the variability given in the numerical predictions was due to sample variability. In investigating the characteristics of those ten instructors, it is discovered that five have never taught statistics, and five have taught statistics. Similarly, inspecting 12 of those instructors whose justifications emphasized the expected value reveals that six had taught statistics and six had not. It is evident that there is some reason for this split, which keeps appearing in all sections of this study, but it is not clear what that reason is.

The concept of repeated samples was also investigated in the interview (Instrument B) where instructors were able to express freely what came to their

minds while solving a problem. There were 12 instructors who took part in the interview. Seven of the instructors had experience teaching statistics and five did not. There were two questions on the survey that investigated instructors' conceptions of variation in repeated samples, B.Q1 and B.Q2.

**Analysis of Question B.1**

Question B.1 (Figure 4.3) presented instructors with four empirical sampling distributions and asked them to identify the fake graphs from the real ones. The question did not clarify how many graphs were 'fake' and how many were 'real'.

The results indicate that instructors had great difficulty assessing these graphs and tended to focus on extreme values and center to aid them in their decision, which led many of them to the wrong conclusion.

Instructors' responses were coded first for the number of correct graphs identified (Graph 1 and Graph 3 were fake, while Graph 2 and Graph 4 were real). However, more importantly, the results were analyzed for the type of reasoning used for such a decision. In particular, it was observed what aspects of the distribution that instructors concentrated on. The results indicate that none of the instructors identified all of the four graphs correctly (Table 4.9) and only one instructor recognized three graphs correctly. The largest percentage identified only two of the graphs correctly. Note that when the results are segmented by instructors' statistics experience, a large percentage of those who teach statistics did not identify any graphs correctly.

**1]** Real/ Fake

A class conducted an experiment, pulling 50 samples of 10 candies from a jar with 750 red and 250 yellow, 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. Which graphs do you think are real? Which graphs do you think are made-up. Explain

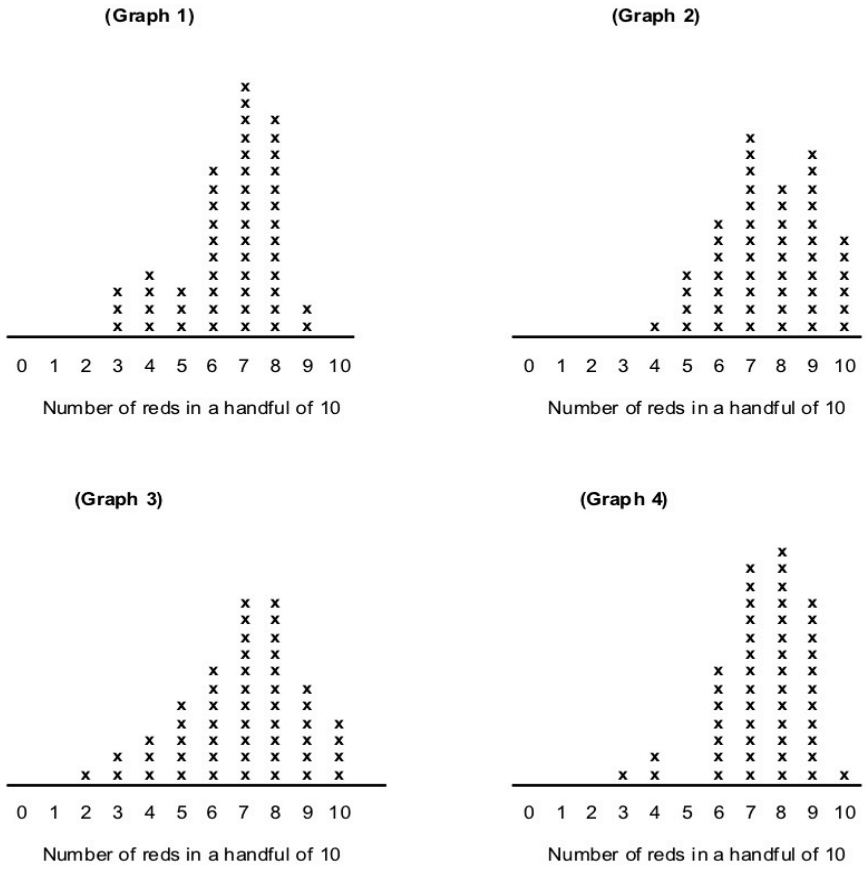



*Figure 4.3.* Instrument B Question 1. Adapted from "Students' attention to variability when comparing distributions" by M. J. Shaughnessy, M. Ciancetta, K. Best, and D. Canada, 2004, in paper presented at the 28th Conference of the International Group for the Pyschology of Mathematics Education, Bergen, Norway.

Table 4.9

*Instructor's Responses to B.Q1*

| Number of correct Identifications* | MT/ST | M | All |
|---|---|---|---|
| | n=7 | n=5 | n=12 |
| 0 | 43% | 0% | 25% |
| 1 | 14% | 40% | 25% |
| 2 | 43% | 40% | 42% |
| 3 | 0% | 20% | 8% |
| 4 | 0% | 0% | 0% |

*Note*. Results based on short answer responses. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics. In order to receive a score of four correct identifications, an instructor would have to had correctly identified Graph 1 and Graph 3 as made up and Graph 2 and Graph 4 as real. Scale adapted from "Students' attention to variability when comparing distributions", by M. J. Shaughnessy, M. Ciancetta, K. Best, and D. Canada, (2004). Paper presented at the 82nd Annual Meeting of the National Council of Teachers of Mathematics.

More relevant than discovering whether instructors' answers were correct or incorrect was investigating the conception of variation utilized by the instructors to make their predictions. Overall, the majority of the instructors did not attend to all aspects of the distributions. Instructors used some of the characteristics of the distribution to make their decisions but tended to focus on only one aspect, for example, extreme values, shape or center, but then did not integrate them to form a firm decision.

**Focus on center.**

Of the 12 instructors, 25% (n=3) analyzed all the graphs according to how accurately a graph represented the average 7.5 red candies, but were confused when this technique proved insufficient.

> *M15: Let me look at this one (pointing to 3). So here is 7.5, I don't know, I will not just count as being fake say here they got 4 times all reds and 7 times got 9 out of 10 red. There were few here, 2 red, 3 red. 4 red. I don't know, I will not quickly say it's fake. I will not accuse them of making it up. I will probably not accuse anyone of making it up.*

> *M26: And then, and then also looking at the means here. Some of these…these means should be fairly close to…to two-thirds. The… so that mean right there… I mean it does… seems to be a little weighted over here. I would also say this mean…umm this means pretty close to two-thirds but that seems a little bit high… so and also there is… the scatter of the mean is probably not that great on this one so you would expect pretty much the… clustering with a…with a mean that is pretty close to two-thirds…so this one looks a little high. I mean so based on mean alone… that's probably about… that's probably about right.*

Note that all of the graphs had a concentration near the average of 7.5. These instructors did not have at their disposal other aspects of a distribution to use as a guide. M15 tried to look at the number of 10 red candies, but was very uncertain making predictions based on that.

**Focus on extremes.**

Of the 12 instructors, 58% (n=7) focused on the extreme; they were particularly disconcerted when they saw many outcomes with 10 candies (i.e. six times 10 red candies in Graph 2). They did not connect the empirical sampling

distribution with the population, which had 75% red candies. This means that 10 red candies are more likely than 0 or 2 red candies. They made their prediction solely on this aspect (number of 10 reds) disregarding aspects of center and spread, and ignoring the distribution of the population. This kind of thinking guided them to the wrong assessment (Table 4.10).

Table 4.10

*Examples of Instructors' Responses to B.Q1 That Focused on Extremes*

| ID | Instructors' responses |
|---|---|
| M19 | This (pointing to Graph 2) has a lot of 10s which makes me very suspicious and I will guess fake. |
| M14 | What bother me here is these tens (pointing to Graph 2). |
| MT13 | If there is one fake or one made up I would probably guess graph 2 intuitively [fake] I don't think that they would be 6 10's that seems highly unlikely to me and then so I can see one or two 10's but intuitively I don't think getting six 10's out of 50 samples seems real. |
| MT4 | (Graph 2) is probably rare because we would expect 7, 9 or a 10 to be rare and you expect some dispersion some variation but that seems to contrive. I never expected to be almost as many 9 out of 10, or 7 out of 10. |

**Focus on shape.**

Samples of the justifications that focused on shape are shown in Table 4.11. Of the 12 instructors, 42% (n=5) focused on the shape. They particularly looked at the regularity of Graph 3 and identified it to be too symmetric or too normal, indicating that the student was cheating. This implies that instructors recognized that the distribution of the repeated samples would have more variation and therefore a perfect graph indicated some kind of cheating. All but one of the five instructors who mentioned shape correctly identified Graph 3. The shape was also considered for the

gap in Graph 4 (no five red candies) and instructors realized that students would not have left the gap. This indicates a high awareness of student behavior and demonstrates an understanding of empirical distribution even if they did not verbalize it.

Table 4.11

*Examples of Instructors' Responses to B.Q1. Focus on Shape*

| ID | Instructors' responses |
|---|---|
| M19 | I will say that this one (pointing to 4) is real because a cheater will never leave an empty space. This one (pointing to 3) is almost too symmetric which make me think that is fake. Too neat. |
| M26 | (Graph 3) does look a little cooked up because it's so symmetric and you might expect a little bit more scatter than that which is fifty samples. |
| MT13 | I am thinking definitely graph 4 is real because if students were to make up numbers I would expect to see a lot of 7,8, 6,7, 8, 9, which there are, but they would not think to make up real numbers, they would definitely have something for 5 before they go down to 4 and 3 so I would think 4 is actually real. Graph 3 looks real but it may be almost too good. |
| M16 | I am pretty sure that 3 is made up because is very even. I don't expect to see that much of a perfect bell curve about 7 [and] 1/2 especially since it is even number in 7 and 8, like if they knew that the teacher would expect 7 [and] 1/2. |
| S12 | I think, I think Graph 3 cheated cause it looks too perfect. It's too good to be true. |

**Focus on spread.**

Out of the 12 instructors, 17% (n=2) focused on the spread. One instructor (ST17) made reference to the sample space instead of the expected spread of the distribution, leading S17 to the wrong conclusion. The other instructor who considered spread was ST3 who used formulas for the binomial distribution to find

the average and the standard deviation and calculated the center plus two deviations to the right and to the left of the center. ST3 found that the expected value was 7.5 then added two deviations (about 1.5) to both sides of the center. ST3 then came up with a reasonable spread for the distribution, from 4.5 to 10.5.

*ST13: So that would run me over the top, so this one 'Graph 3' is real goes up to 10 is pretty significant to get a 10, this is good.*
*[Pointing to graph 4] this one looks like it is cheating,*
*[Pointing to first one] I think this one is probably also cheating because it does not seem up enough, and drops off too quickly.*
*[Pointing to graph 2] I think this is probably real because I don't think that if the students made it up it would not have dropped at 8 and pop up again at 9.*

Note that the calculations brought ST13 the surprise result that 10.5 was out of the range of the provided distribution, and he/she then correctly identified Graph 2 as cheating because of the spread of the data. However, ST13 incorrectly identified Graph 3 using the same argument. This indicates that he/she did not consider the regularity of the shape of the graph to be a surprise. Therefore, the two instructors who considered spread were not guided by this knowledge to arrive at the correct decisions. ST13 identified only two graphs correctly and ST17 identified none of the graphs correctly.

From the analysis of this question it can be deduced that some of the instructors found key principles to identify which graphs were real and which ones were fake. In some cases, the awareness of students' thinking helped them recognize graphs correctly. In other cases, recognizing the randomness of the empirical sampling distribution helped them identify graphs whose shapes were too neat. Recognizing the spread of the data helped them to identify graphs that did not have

the desired spread. However, none of the instructors utilized all these elements in making their decision and therefore they were unable to identify all of the graphs correctly. Also, too much emphasis on center and extreme values led instructors to the wrong conclusions. It seems that a more integrated approach is needed to arrive at the appropriate conclusions.

**Analysis of B.Q2**

Question B2 (Figure 4.4) was also designed to unveil instructors' conceptions of variation when predicting repeated samples. This question had two parts, 'a' and 'b'. In part 'a' instructors were asked to predict six repeated samples, while in part 'b' they were asked to choose from a list. The results indicate that the majority of the instructors predicted values that showed consideration for variability, while some insisted on the expected values and showed no consideration for variability.

The results were coded for part 'a' similarly to A.Q1c. It was first considered whether the answer demonstrated variability (different numbers for the six samples) or not (all numbers were the same). Then the responses showing variability were coded as either appropriate or inappropriate. Following Reading and Shaughnessy (2004), inappropriate responses were those that were too low (M25: 2,3,2,1,3,5), too high (M43: 5,6,7,8,9,10) or too broad (M52: 0 to 10). Appropriate responses were (M34: 5,5,6,7,7,8), or (M28: 3,4,5,5,6,7) or (MT: 5,6,6,7,7,7). Appropriate responses were those predictions that were not too low, or too high, nor too broad.

**2]** A bowl has 100 wrapped hard candies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Jenny pulls out a handful of 10 candies while blindfolded, counts the number of reds, and tells her teacher. The teacher writes the number of red candies on a list. Then, Jenny puts the candies back into the bowl, and mixes them all up again. Five of Jenny's classmates, Jack, Julie, Jason, Jane and Jerry do the same thing. They each pick ten candies count the reds, and the teacher writes down the number of reds. Then they put the candies back and mix them up again each time.

a) What do you think the teacher's list for the number of reds is likely to be? Explain why you chose those numbers.

b) If you were asked to choose a response to this question from the following list, circle the one that you would choose. Explain why you chose that one.

A) 5, 9, 7, 6, 8, 7  B) 3, 7, 5, 8, 5, 4  C) 5, 5, 5, 5, 5, 5  D) 2, 4, 3, 4, 3, 4  E) 3, 0, 9, 2, 8, 5

*Figure 4.4.* Instrument B Question 2. Adapted from "Reasoning about variation: Student Voice" by C. Reading & J. Reid, 2007, *International Electronic Journal of Mathematics Education*, 2(3), pp. 110-127.

The results indicate that 67% (n=8) of the instructors predicted variability in the six repeated samples, while 25% (n=3) did not write a list of values and 8% (n=1) wrote two responses: a list with all same values (5,5,5,5,5,5) and a list with different values (4,5,5,5,5,6). Within the instructors who predicted variability, the results indicated that all predicted reasonable amounts of variability in the repeated six samples. Also, the analysis of the responses for part 'b' indicates that all but one instructor chose the appropriate list of six repeated samples.

Within the justifications for part 'a', it was found that 67% (n=8) gave reasonable predictions justifying that the numbers were close to the expected, while 17% (n=2) of instructors did not provide a justification for their predictions, and

17% (n=2) gave justifications that relied solely on the expected value. For part 'b', the instructors' justifications reflect that most of the instructors (75%, n=9) had an adequate awareness of variability; they chose the appropriate list by a process of elimination. In the process they revealed all the aspects of the distribution that they were considering. So this question seems very revealing of instructors' thinking. A typical response for part 'b' was:

*ST3: Not that one [C: 555555] because for sure it will not happen all that way.*
*[D: 2,4,3,4,3,4] this is very close but they are kind of short. I don't think is this one, skewed. I don't know which way is skewed but it is too left sided.*
*[B: 3,7,5,8,5,4] 5 and 5, 4, and it could be, a six but there is not, but there are two under and two over that is a possibility. [A: 5,9,7,6,8,7] this is too high*
*[E: 3,0,9,2,8,5] this is too extreme. So I think this one B is the best*
*So why did I chose that one, process of elimination.*

It becomes evident in this type of explanation that instructors considered not only the center of the data but also the reasonable spread around the center. This demonstrates a high awareness of the idea of variability, which was not that evident in the previous questions (A.1, A.2, A.3, and B1) dealing with repeated samples. It seems that the list format in question B.2 triggered the appropriate knowledge. This suggests that the questions themselves seemed to be a big influence in the way they unveiled instructors' conceptions of variation. This is interesting since it may be the same for students too and warrants further investigation.

**Summary of Results for Research Question One**

From the analysis of questions A.1, A.2 and A.3 that deal with repeated samples, several aspects of instructors' considerations of variation appeared. Firstly,

these questions showed some tension in instructors' thinking and reasoning as expressed by contrasting the numerical responses with the justifications. Secondly, throughout the questions in this section it became evident that there is a tendency to focus on the center of the distribution or the expected values. Thirdly, instructors who predicted variability seem to lack consistency across the questions in this section. Fourthly, instructors' education and teaching experience do not seem to relate to the type of responses instructors gave. Finally, besides understanding instructors' conceptions of variations, these questions (A.1, A.2 and A.3) unveiled the fact that numerical responses are not necessarily a good indicator of the instructors' conceptions of variation and highlight how the question format and context seem to influence the type of response.

It became evident when analyzing these questions that the numerical predictions and justifications did not necessarily concord. Some instructors did not predict variability in the written repeated samples, but then clarified in their justifications that variability would occur. For example, M33 wrote the expected values (10,10,10,10,10,10) but then said, "it won't really be 10-10-10... but I don't have any way of telling which numbers will come up more often than others. I know they will come up different numbers of times, somewhere between 5 and 15, say, centered around 10". In some cases, instructors mentioned that they could not judge whether the question asked them to consider theory or practice (M18). On the other hand, some instructors predicted variability in the repeated samples but their justifications did not clearly reveal their conceptions of variation. For example,

MT46 predicted (9,10,12,11,10,8) and stated, "Each number is equally likely, so the counts should be about the same". This tension permeates the questions analyzed here.

A pattern emerged among instructors who did not predict variability. They tended to provide numerical values as well as justifications that concentrated mostly on the center of the distribution, thus ignoring other aspects of the distributions, or they tended to focus on the expected value and ignored sample variability.

On the other hand, when considering consistency in the responses, the majority of instructors seem to have, at best, weak consideration of variation. This comes to light when looking at the overall pattern of responses for these three questions. First, only one instructor (M11) predicted variability consistently in all the three questions, and his/her justifications also demonstrated an awareness of sample variability when taking repeated samples. This means that 98% (n=51) of the instructors were inconsistent throughout the items in acknowledging variability in repeated samples. However, it is important to note that only 15% (n=8) of the instructors did not predict variability in any of the sections of these three questions. This means that 85% (n=44), at one point or another, considered variability in repeated samples. It remains unclear what triggers instructors to recognize variability in repeated samples in some questions and not in others.

One fascinating result that emerged from the questions dealing with repeated samples is that education and statistics teaching experience do not seem to be related to the type of responses given to these questions. Also, as shown in the preceding

results, the percentages do not differ greatly when the questions were analyzed by instructors' characteristics. One unexpected result with regards to instructors' characteristics is that the only instructor who seemed to have a consistent and accurate conception of variation (M11) does not have a degree in statistics and has never taught statistics. Another result that may be considered unusual, for example, was the responses from ST17, who has a degree in statistics and teaches statistics often, and who was one of the few instructors who did not predict variability in any of the questions presented here.

These questions also helped in demonstrating the importance of looking at both types of responses, i.e. short answer type of responses as well as the justifications behind these short answers. This connection held the key, in some cases, to understanding instructors' thinking and brought awareness of the types of conflicts that some instructors experienced with these questions. Overall, these three questions dealing with repeated samples represent the starting point of understanding not only the conceptions of variability that instructors have, but they also bring to light several questions that are a springboard for further research.

From the analysis of questions B.1 and B.2 that deal with repeated samples, several new understandings of instructors' considerations of variation were revealed. In question B.1, dealing with empirical sampling distributions, instructors seemed to struggle to adequately judge which graph was real and which graph was fake. Only one of the 12 identified all four graphs correctly and the majority did not identify any of the graphs correctly. One possible explanation is their lack of exposure to such

distributions and another is their compartmentalized usage of aspects of a distribution. Many focused only on the extreme values, while others focused only on the shape or center. Only one of the instructors demonstrated use of center and spread for the decision, but still failed to identify the appropriate graphs.

Question B.2b proved to be a crucial type of question in this study because it revealed the depth of instructors' thinking regarding variability in the repeated samples. All of the other questions discussed in this chapter did not produce responses that highlighted instructors' conceptions of variability. On the contrary, from the prior questions it was deduced that instructors show little consideration for variability in repeated samples. While it is important to take into consideration that question B.2b was only given to the 12 instructors who did the interview, it seems to provide insight into instructors' thinking of repeated samples. It may be necessary to investigate seriously the effect that the type of question has on the response that instructors, and students for that matter, provide. This is important if the goal is to understand how students and instructors think about variability.

# Chapter V

## Results Pertaining to Research Question Two

Several questions were designed to unveil instructors' conceptions of variation when dealing with one data set. There was a total of five questions: A.5 and A.9 from the survey and questions B.3, B.4 and B.5 from the interview, which were designed to answer Research Question Two, namely:

What statistical concepts do instructors utilize to describe and/or decide the best way to represent one data set and does the context of the problems guide those decisions?

- in the presence of unusual variation?
- in the presence of several graphical displays?
- in a histogram?
- How are teaching experience and education related to instructors' responses?

Questions A.5 and A.9 were answered by the 52 instructors who took the survey (Instrument A), and questions B.3, B.4 and B.5 were asked only of the 12 instructors who took part in the interview (Instrument B). The main purpose of questions A.5, A.9, B.3, B.4 and B.5 was to gain insight into instructors' considerations when dealing with one data set. Since it is recognized by scholars that "the role of context is critical and the importance of context is among the things that distinguish statistics from mathematics" (Peck, 2005, p. 1), one of the goals was also

to investigate how instructors decide the best way to represent a data set and how the context of the problem guided instructors' decisions.

**Analysis of Question A.9**

Question A.9 was designed to discover how instructors decide what the best measure of center is when they encounter unusual variation in a data set. Question A.9 (Figure 5.1), provided instructors with nine data points representing the weight of a small object that students have measured. The value 15.3 in the data set was an obvious outlier. Instructors were asked to determine what the most appropriate method would be if the goal was to find the actual weight of the object. As seen in Figure 5.1, this item had four possible choices for instructors to pick: a) Method I: Use the most common number, 6.2; b) Method II: Use 6.15 since it is the most accurate weighing; c) Method III: Use the result of adding up the 9 numbers and dividing by 9; d) Other. This question also asked instructors to justify their answer.

The results showed that the majority of the instructors recognized the distorting effect that the outlier would have in estimating the weight of the object and chose to discard it and then average the rest, take the median or the mode. However, several instructors decided that adding all of the numbers and dividing them by nine was the best measure even in the presence of the outlier.

The results were first analyzed by looking at the percentage of instructors selecting each of the four choices. The results indicate that Method I was chosen by 29% (n=15) of the instructors, Method II by 4% (n=2) of the instructors, Method III

by 23% (n=12) and 'Other' by 44% (n=23) of the instructors. These percentages indicate that 77% (n=40) of all instructors recognized the effect of the outlier and used alternate methods.

**9]** Nine students in a science class weighed a small object on the same scales separately. The weights (in grams) recorded by each student are shown below.

6.2 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.2

The students want to determine as accurately as they can the actual weight of this object. They may use the following methods:

I. Use the most common number, which is 6.2

II. Use the 6.15 since it is the most accurate weighing.

III. Use the result of adding up the 9 numbers and dividing by 9.

As teacher, what method would you prefer your students use?

a) Method I

b) Method II

c) Method III

d) Other_____

Explain your choice.

*Figure 5.1.* Instrument A Question 9. Adapted from "The measurement of school students' understanding of statistical variation" by J. M. Watson, B. A. Kelly, R. A. Callingham, and J. M. Shaughnessy, 2003, *International Journal of Mathematical Education in Science and Technology,* 34(1), pp.1-29.

However, it is interesting to note that 23% (n=12) of instructors chose Method III, which consisted of adding all nine values (including the outlier) and dividing by nine to find the accurate weight of the object. This means that the weight of the object would be about seven grams, which is higher than most of the values in the data. It is not clear why instructors chose Method III as an appropriate way to represent the weight of the object. It may be theorized that the instructors did not realize that the outlier was there, but in the justifications they gave it was found that they did indeed realize that the outlier was there, but still decided that adding all the numbers and dividing by nine was the best measure.

Table 5.1 shows the justifications for the instructors who chose Method III. It can be deduced that, in spite of the presence of the outlier, they considered the mean to be the best representative. Some of the explanations were puzzling as some instructors brought information that does not pertain to this particular scenario, for example " it is likely that the same error was made every time", or "let's assume the measurements are symmetrically distributed". It is not clear what prompted these kinds of responses, but it can be guessed that instructors were using theoretical knowledge to answer the question instead of paying attention to the context of the data and to what the task at hand required.

It may be argued that these 12 instructors did not have much experience dealing with outliers, as they may not have taught statistics. Table 5.2, however, shows that six out of the 12 regularly teach statistics and two of the 12 had a degree in statistics. It is not clear why they think that Method III is the best measure in the

presence of an outlier, particularly when Method III would produce an average that is larger than the majority of the data points.

Table 5.1.

*Examples of Instructors' Justifications for Choosing Method III in A.Q9*

| ID | Instructors' Justification |
|---|---|
| ST1 | Let's assume that the measurements are symmetrically distributed around the true values, + that SD is finite. Then x-bar is a reasonable estimate of mu (if the SD (i.e. variance is infinite then x-bar is no better than any other measurement) |
| ST3 | There is variability in measurement. The mean would be a good way to take that into account. |
| MT5 | This is how to tell them to find their av. On Hmw & tests taken in class |
| MT9 | For an unbiased scale, the variation is likely the result of errors in how the students measured the results. Unless I have a reason to suspect that it is biased, the mean is the best way to balance out those errors. |
| M20 | The arithmetic mean is the most commonly used, and therefore most familiar, measure of tendency. |
| MT24 | This method takes into account the entire data set and gives the mean value. |
| MT28 | If there was an error(s) making the measurements, it is likely that the same error was made every time so averaging the observations seems like a good idea. |
| MT29 | I'm a math teacher - - I'd probably break the scale. |
| M36 | not a big spread in the data. |
| M50 | This method takes an average; it best eliminates discrepancies in weight due to experimental error. |

*Note.* Method III: Use the result of adding up the 9 numbers and dividing by 9. Two instructors did not justify why they used Method III

Within the justifications of instructors who chose 'Other' (n=23), it is found that they reasoned about the effect of the outlier and recognized that the outlier would skew the measures and therefore they decided to either take out the outlier and average the rest, or to take the median of the values since the median is not affected by outliers, as observed by MT13 "I would disregard the outlier 15.3 + find the mean of the other 8 #s. Or. I'd rather use the median of the 9#s".

Table 5.2 shows that there are no major differences in the percentages of the people in each category who chose a particular method. When the justifications are looked at under the light of instructors' teaching experience (Table 5.3), it is discovered that the percentage of the entire group is similar to the percentage of instructors who teach statistics (MT/ST) and the group of instructors who do not teach statistics (M). This suggests that statistics teaching experience does not seem to

Table 5.2

*Instructors' Chosen Method in A.Q9 by Instructor Group*

| | M<br>n=23 | MT<br>n=22 | ST<br>n=7 | All<br>N=52 |
|---|---|---|---|---|
| Method I | 26% | 32% | 29% | 29% |
| Method II | 4% | 5% | 0% | 4% |
| Method III | 18% | 27% | 29% | 23% |
| Other | 52% | 36% | 42% | 44% |

*Note*. Based on Short Answer Responses. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

produce a different type of response.

Table 5.3

*Instructors' Justification Most Commonly Expressed in A.Q9*

| | M<br>n=23 | MT/ST<br>n=29 | All<br>n=52 |
|---|---|---|---|
| Because of outlier best to use: mean without outlier, median or mode | 48% (11) | 62% (18) | 56% (29) |
| Other | 30% (7) | 38% (11) | 34% (18) |
| No Justification | 22% (5) | 0% (0) | 10% (5) |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

Overall this question brought to light the different ways instructors deal with unusual variation. The majority recognized that in the presence of an outlier, the mean without the outlier, the median or the mode would be the best way to represent the data, and while there may be arguments about which measure of center is best, recognizing the presence of the outlier and knowing the alternatives was the main point of this problem. However, several instructors, regardless of their statistics teaching experience, considered that adding all the numbers (including the outlier) and then dividing by the total number was the best measure for this data. The logic behind this kind of argument, even though they expressed their reasoning, did not follow the most common method of dealing with outliers.

**Analysis of Question A.5**

Question A.5 (Figure 5.2) had one data set representing the braking power of a car, which is determined by the stopping distance of a car going from 40 mph to 0 mph. The measurements were in inches and the data were displayed in a list, and also in three different graphical representations. The first graph showed only occurring values, eliminating values with zero frequencies, the second showed all occurring values and included the zero frequencies, and the third showed data in intervals and presented data as aggregate. This question had four sections (a, b, c, and d) and each section had a short answer response and a justification.

While the short answer responses provided a general pattern of response, the justifications revealed the reasoning behind such responses. Table 5.4 shows the results for the short answers of all the parts of this question. It can be seen that there was consensus in the chosen responses except for part b.

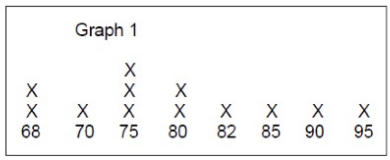
For example, in part 'a' almost all instructors (94%, n=49) stated that the three graphs differ in the way they showed the braking power; for part 'b', the majority (73%, n=38) stated that one graph showed more variability than the others; for part 'c' the majority (88%, n=46) stated that Graph 3 would be the best graph to show that the car was fairly consistent; in part 'd' the most common response (46%, n=24) was to decide that Graph 2 was the best graph to reach a conclusion. Another salient characteristic of the short answer responses was that several instructors (21%, n=11) responded with a yes and no to part 'b' of the question. They stated 'yes' if variability was based on the graphical display and 'no' if the data were considered

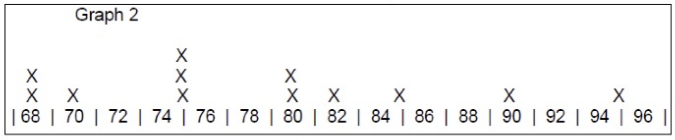
for variability. Additionally, in part d several instructors (15%, n=8) did not understand the question and instead of giving a response they asked: "What conclusion?"

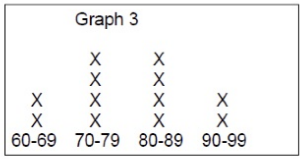
**5]** A new car was being tested to see how well the brakes worked. The test engineer measured how many inches the car took to slow from 40 mph to 0 mph; the fewer inches taken, the better the braking power. Twelve trials were run, under the same road conditions and with the same test driver. Here were the results ( to the nearest inch):

| Stopping Distance (in) | | | |
|---|---|---|---|
| 68 | 68 | 70 | 75 |
| 75 | 75 | 80 | 80 |
| 82 | 85 | 90 | 95 |

The engineer was then trying to decide how to graph the results. She came up with the following three graphs for representing the data:




a) Do these graphs differ in the way they show the braking power? If so, how?

b) Do you think one graph shows more variability in the results than others? Explain.

c) If the engineer wanted to suggest that the car was fairly consistent in its braking power, which graph would you suggest she use, and why?

d) Does one graph help you more than the other in making your conclusion?

*Figure 5.2.* Instrument A Question 5. Adapted from "Elementary preservice teachers' conceptions of variation" by D. Canada, 2004, Portland State University, Portland.

When the short answers responses were investigated in the light of instructors' education and teaching statistics experience (Table 5.4), it became evident that the percentage of responses for each section of this question was similar for all of the groups. Note that instructors who have a degree in statistics mostly answered "yes and no" which the researcher considered equivalent to "yes". This indicates that statistics teaching experience does not seem to produce a different type of response.

Table 5.4

*Instructors' Responses to A.Q5*

| | | All n=52 | M n=23 | MT n=22 | ST n=7 |
|---|---|---|---|---|---|
| Part a | Yes | 94% | 92% | 100% | 86% |
| | No | 4% | 4% | 0% | 14% |
| | No Response | 2% | 4% | 0% | 0% |
| Part b | Yes | 73% | 78% | 82% | 29% |
| | No | 4% | 4% | 0% | 14% |
| | Yes and No | 21% | 14% | 18% | 57% |
| | No Response | 2% | 4% | 0% | 0% |
| Part c | Graph1 | 2% | 0% | 5% | 0% |
| | Graph 2 | 8% | 9% | 9% | 0% |
| | Graph 3 | 88% | 87% | 86% | 100% |
| | No Response | 2% | 4% | 0% | 0% |
| Part d | Yes Graph 1 | 6% | 4% | 9% | 0% |
| | Yes Graph 2 | 46% | 48% | 45% | 44% |
| | Yes Graph 3 | 6% | 4% | 5% | 14% |
| | No | 10% | 10% | 5% | 28% |
| | What Conclusion? | 15% | 17% | 18% | 0% |
| | Other | 17% | 17% | 18% | 14% |

*Note.* Based on Short Answer Responses. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

While there are many way to analyze the justifications of these sections (a, b, c, and d), the researcher looked at the responses for evidence of four important characteristics that represent not only understanding of variability but also the importance that context has in statistics tasks. The first characteristic examined was the consideration of the context. According to Roxy Peck (2005), "in statistics, meaning comes from context, and it is the interpretation of the analysis in context that is the ultimate desired outcome of analyzing data." (p. 1). Therefore instructors' justifications were analyzed by their consideration of the context in all four parts of this question.

The second characteristic was the consideration of data as an aggregate. According to Konold and Higgins (2002), students tend to progress from seeing the data as a collection of unique individuals to the idea of seeing the data as an aggregate where emergent features such as center, spread and shape become more evident (p. 179). Therefore, instructors' responses were analyzed for their ability to recognize the benefits of data as an aggregate.

The third characteristic was the consideration of zero frequencies (non-occurring data values). Konold and Higgins (2002) explained that the decision of including zero frequencies depends on the question being investigated; they also stated that whether or not to include zero frequency "can drastically affect how they perceive those displays" (p. 182). Therefore, instructors' responses were investigated for the recognition of missing values as zero frequencies, as represented in Graph 1.

The fourth characteristic was the consideration of variability in data.

According to Shaughnessy et al. (2004), variability is understood as extremes or possible outliers, as spread, as the heights of the columns in stacked dot plots, as the shape of the dispersion around the center, and as distance or difference from expectations (p. 29). Therefore, instructors' responses were investigated for descriptions of variability in parts b and c.

**Context.**

When instructors' justification were analyzed for reference to context, it was found that only 23% (n=12) gave even a small indication of the context and even then they did not use the context to guide their decisions. The other 77% did not mention context at all. Those who acknowledged context mostly wrote about the variable of interest, namely, the braking distance and some mentioned the unit (inches). However, instructors did not make their decisions by paying attention to the context as indicated in their justifications. Only one of the 52 instructors wrote in response to part 'd', "fair? Consistent? 28 inches for 40 mph seem great. Consistent" (M4). This instructor did not use this information to answer any of the other parts of the question where this kind of awareness would have been useful. However, M4 recognized that the spread was small, that the unit was inches, and that therefore the data had very small variability, indicating that context matters.

**Data as an aggregate.**

The benefits of seeing data as an aggregate were not evident in instructors' justifications. The results indicate that only 10% (n=5) recognized the usefulness of Graph 3 as a graphical representation that was helpful to understand the overall

behavior of the car braking distance. For example, MT40 wrote, "Yes. #1 has an uneven horizontal scale, and doesn't show the true spread. #2 gives too much emphasis to individual items. Graph #3 does the best job in showing the shape of the distribution." This means that 90% (n=47) of the instructors did not see the benefit of displaying the data as in Graph 3. On the contrary, statements like it "loses info", or "masks variability", or "best way to cheat" were common. Table 5.5 shows examples of instructors' exact statements. One can conclude that instead of seeing aggregate data as a way to see the bigger picture and understand what is happening in general, instructors tended to believe that Graph 3 loses information or concealed some information purposefully.

Table 5.5

*Examples of Instructors' Considerations for Graph 3 in A.Q5c that Suggest it Hides Information*

| ID | Instructors' Written Responses |
|---|---|
| M23 | Probably #3 in this case, because the groupings suppress the variability a bit |
| M27 | Graph 3. Because most people are visual, and the effect of this particular representation is the most persuasive to her cause |
| M35 | To achieve that "lie with statistics", they should use #3 |
| ST2 | 3 would mask some of the actual variability |
| M42 | Maybe graph 3, but that graph "hides" some of the data. |
| MT46 | Graph 3. It does a good job of hiding the variation in the results. |
| M45 | 3, because it shows less detail and consequently less variability |
| M47 | Graph 3. The graph suggests that there is very little variability |
| ST48 | Graph three- if he wanted to deceive his audience, because most people do not look at the scale, just at the height of the bars |

**Zero frequencies.**

The results after analyzing instructors' comments about the missing values of Graph 1 revealed that 56% (n=29) did not mention the scale at all, while 44% (n=23) of the instructors' described the scale as missing values. Of the 23 instructors, five stated that the wrong scale made the graph misleading. For example, M35 wrote, "Yes. #1 is confusing due to the unequal steps between the horizontal axis values; misleading!" None of the 52 instructors mentioned that the problem with the scale was due to the zero frequencies not being included in the graph; they recognized the error but did not mention the source of the error.

**Considerations for variability.**

The analysis of instructors' justifications about variability (part b) revealed that 94% (n=49) of instructors stated that Graph 2 showed more variability. While not everyone gave the same reasoning, the largest percentage (48%, n=25) of instructors wrote that Graph 2 showed more variability due to the spacing of individual data values and because it clearly showed the extreme values. While this point of view is plausible, it suggests that instructors focused mainly on the distance between individual values instead of the distance from the center. Moreover, the lack of consideration for context made values look very spread out; however, when the context is considered, namely distance in inches as a measure of the braking power of a car going from 40 mph to 0 mph, it becomes more clear that the data points were only a few inches apart.

Another interesting finding in the justifications about variability was that none of the instructors identified Graph 3 as a way to visualize variability; on the contrary, 8% (n=4) stated that it was hard to measure variability in Graph 3. This suggests that these instructors did not consider the display of data as an aggregate as a way to visualize variability. Another type of justification about variability that appeared in instructors' responses revealed that 13% (n=7) of the instructors indicated that numerical measure of variability was one and the same, but visually graphs seemed to show more variability. For example, MT21 wrote, "the standard deviation does not change, based on the original data. However, different data classes can result in a graph that looks to have more variability". Regardless of the type of response instructors gave, it seems that their judgment of variability was not based on the context of the data and it was not viewed as a measure of spread from center.

While the instructors' justifications have been categorized into four main parts, the results indicate that the percentages within each category are similar even when instructors' teaching experience is considered (Table 5.6). It appears that teaching of statistics does not seem to be related to the type of justifications instructors gave.

Table 5.6

*Instructors' Justifications for A.Q5 by Group*

| Categories | Levels | M n=23 | MT/ ST n=29 | All n=52 |
|---|---|---|---|---|
| Context | Yes | 26% | 21% | 23% |
| | No | 74% | 79% | 77% |
| Show more Variability | Graph 2 | 74% | 69% | 71% |
| | Other | 18% | 28% | 23% |
| | No difference | 4% | 3% | 4% |
| | No response | 4% | 0% | 2% |
| Mentioned Graph 1 Scale | Yes | 39% | 59% | 50% |
| | No | 61% | 41% | 50% |
| Value of Data as an Aggregate | Yes | 13% | 7% | 10% |
| | No | 87% | 93% | 90% |

*Note*. M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics. The cells for ST and MT have been collapsed because the count of the cells was less than 5.

**Analysis of Question B.5**

Question B.5 was part of the interview and therefore only 12 instructors answered it. This question had three parts (Figure 5.3); the first two parts asked instructors to state what could be deduced from the graphical representations; in other words, what information could be extracted by looking at the graphs. The third question asked instructors to decide which of the two graphs told the story better. Both graphs included identical data values; however, in Graph 1 the scale only included occurring data points, in other words the zero frequencies were skipped on

the scale, while the second plot used all the values and included an even scale that represented zero frequencies.

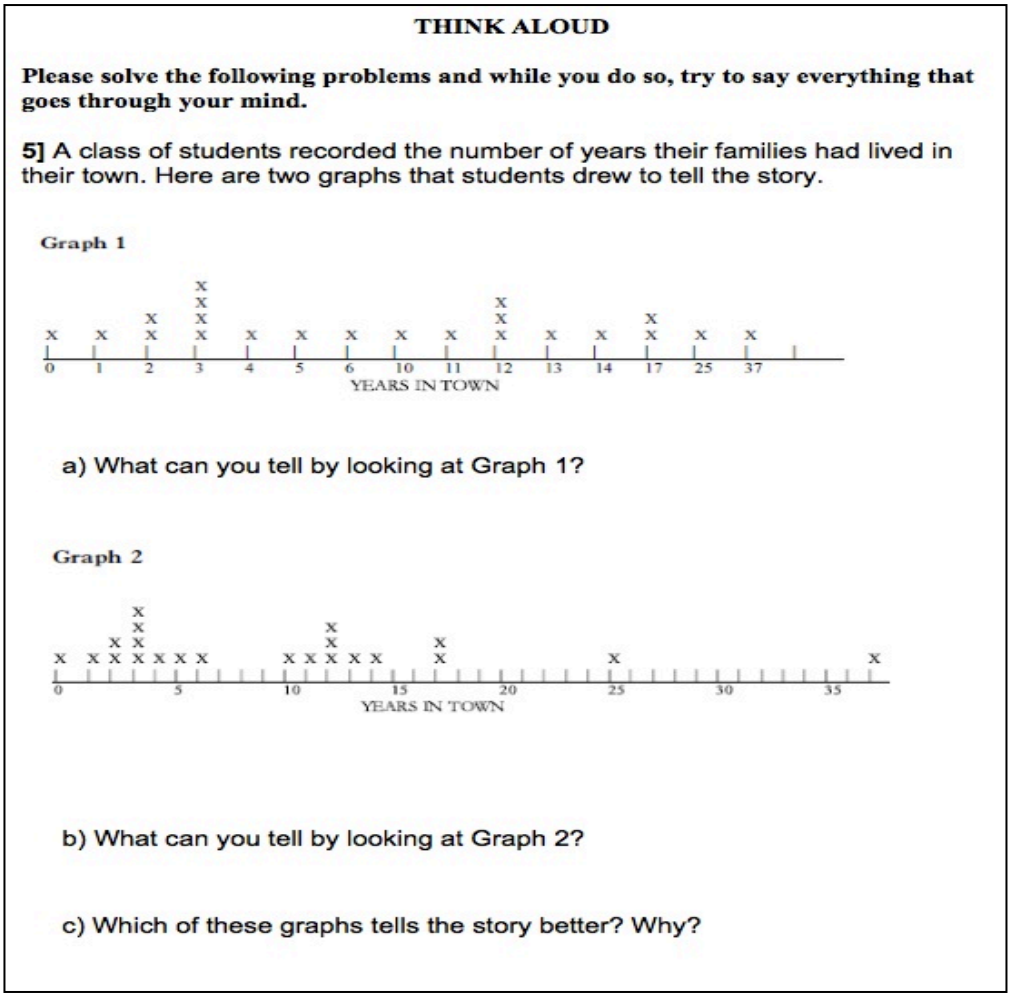



*Figure 5.3.* Instrument B Question 5. Adapted from "The measurement of school students' understanding of statistical variation" by J. M. Watson, B. A. Kelly, R. A. Callingham, and J. M. Shaughnessy, 2003, *International Journal of Mathematical Education in Science and Technology*, 34(1), pp.1-29.

Several important components were considered when analyzing instructors' responses: a) The consideration for context; particularly, the explicit identification of the variable of interest "years in town", b) The ability to clearly recognize that the

missing values on the scale corresponded to the zero frequencies on the data set, c) The ability to recognize the limitations of graphical display that does not include zero frequencies, but still recognize that some data summaries could be extracted from such display, d) The ability to recognize that a better scale permits better description of the overall trend of a distribution, and e) The ability to take context into consideration when making conclusions.

The data were first analyzed by looking at the number of instructors that utilized each of these components. Each instructor response was then analyzed for the number of components utilized in their complete description in all parts of the problem. The results of the number of instructors that utilized each of these components indicate that only 17% (n= 2) of the instructors described the variable of interest. Only 17% (n=2) identified the gaps as the absence of zero frequencies, even though 75% (n=9) recognized the gap in Graph 1. Also, 58% (n=7) recognized how the scale of Graph 2 permitted them to better understand the overall patterns and described the shape of the distribution as well as mentioned the overall trend. All of the instructors (n=12) concluded that Graph 2 was the best display but their justifications were based solely on the scale of the data without mentioning the context.

In order to understand the number of components each instructor utilized in their descriptions, the responses were given codes from zero to three according to the number of components utilized by the instructor. For example, Code 0 was given to

instructors who did not mention any of the components and gave reasoning that was not discernable.

*M14: I don't know. Anybody, don't have the right to call it their town unless they live less than three years or something.*
*MT4: Seems misleading because this stick marks are shown and some they aren't ah.... more mobile. [Written not spoken] no predictability more mobile than not.*

Note that these instructors are not making much reference to the data at hand and gave reasoning that is hard to understand. Only 17% (n=2) of the 12 instructors responses received Code 0. Code 1 was given to instructors who mentioned only one of the five above-mentioned components.

*MT13: well graph one is really not drawn to scale so the numbers are skipped around so makes it look like there is less variability than there really are, and graph 2 is drawn to scale so graph one is a little more deceiving on graph 2 you can tell the shape and the variability much better ah.... so you can tell from graph 2 that the majority of the people in that town have only been in that town for less than 15 years and c graph 2 tell the story better because graph one is more deceiving by not drawn up to scale is more misleading to the person who is reading the graph.*

MT13 did not mention element a, b, c, or e, and only described some aspects of the overall pattern of the distribution in Graphs 2. He/she recognized the gaps, but did not recognize them as zero frequencies; he/she did not utilize context in any statements even to identify the variable of interest. There were 25% (n=3) of the instructors whose responses received Code 1.

Code 2 was given to instructors who mentioned two of the five above-mentioned components.

*ST3: So graph one is a dotted plot there is family that lived there 37 years. Most of them more than anything else 3 years so you can tell that there is quite a variety that some families lived there a long time that there are some families that didn't live there very long and overall except for the fact that the scale is not evenly marked overall it is not that*

*unevenly distributed except for the scale. Okay Graph 2 what can you tell? Is this the same thing? Okay I love this, this is beautiful so by fixing the scale it shows that there are so many more families that have lived there fewer years so this is ... a this is to scale and this is not to scale and all that I will say right now is, wow!! What a big difference it gets. Which of these graphs it tells story better? Graph 2 tells the story much better because it shows you visually even though my brain knew that there were gaps there is nothing from 7, 8, 9 or from 17 to 25 my brain knew that but my eyes did not really really comprehend that yet when I look a graph two since it is evenly spread out I can see that very clearly.*

ST3 did not mention elements b, or e. ST3 described some of the characteristics of the data from Graph 1 and recognized the gaps but did not identify them as zero frequencies. He/she did not explicitly mention the variable of interest, but made reference to it by mentioning the years the family lived in town. Then ST3 recognized how the changes of the scale provided a better understanding of the overall distribution and commented on it, but still did not utilize context for the conclusion. There were 33% (n=4) of the instructors whose responses received Code 2.

Code 3 was given to instructors who mentioned three of the five above-mentioned components.

*MT10: Well it is a bad scale, don't skip, include spaces where the zeros are. The min is zero new to town max is 37 long time in town. I can probably get a 5 number summary quartiles and medians if I wanted it to ah.... the mode is 3, a lot of descriptive statistics you can get. Ah... good scale graph 2 ah...there is a big gap from 25 years in town to 37 years in town and also gap from 17 to 25 and few smaller gaps, ah... skewed right that mean the mean is less than median no more than median the outliers pulls the mean to the... that is skewed right I guess the median (...) is seven I am not going to count. The graph 2 is much better, scale is much better.*

Note that MT10 mentioned components b, c and d, but did not mention components a or e. M10 not only recognized the gaps on the first graph, but also mentioned that the gaps represented the zero frequencies. M10 also recognized that even though there were gaps in the scale some descriptive statistics (like the mode)

could still be drawn; therefore he/she proposed several summaries and mentioned the mode of the data. MT10 also described the overall trend of the data. Out of the 12 instructors who took this interview, only M10 received a Code 3.

Code 4 was given to instructors who mentioned all of the five above-mentioned components.

*ST12: Here are two graphs. Okay so years in town. Yah… all you can tell from graph one is the frequency in each year because you've eliminated all the missing years… the ones that… like you've eliminated all the zeros. So you don't really get the scale… so you can't get the shape… but you can tell that, what you can tell from this graph is that uh three years is the mode. The most umm the most frequent. And in this one you can get that it's, you also can get that the shape that it's more skewed to the right… and one family lived there a long time so you get more information from the second graph. Yah, I see… graph C tells the story better cause it shows that, that uh it's heavily skewed to the right. That there is a couple of families that really are quite a bit different than most of them, what you can even see two small clusters here as well… which are interesting as well. So it looks like a lot of families leave in five years, and then… then they sort of don't leave and they got a bunch around here and then they leave maybe a generation later when the kids finish school or something. And then you got people who live in that town forever.*

Note that ST12 mentioned all the five components; a) The consideration for context in implicitly identifying the variable of interest "years in town", b) The ability to clearly recognize the gaps in the scale and identify them as zero frequencies, c) The ability to recognize that some data summaries could be extracted from Graph 1, but recognize its limitations, d) The ability to acknowledge the benefit of better scale in Graph 2 and use of new display to describe the overall trend of the distribution e) The consideration of context to make conclusions. Out of the 12 instructors who did the interview only 17% (n=2) gave this kind of complete description mentioning all five components.

Table 5.7

*Instructors' Justification Codes for B.Q5*

| | M<br>n=5 | ST/MT<br>n=7 |
|---|---|---|
| Code 0 | 20% | 14% |
| Code 1 | 20% | 29% |
| Code 2 | 40% | 14% |
| Code 3 | 0% | 29% |
| Code 4 | 20% | 14% |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

When the data are parsed according to instructors' statistics teaching experience (Table 5.7) it is discovered that 57% of those who teach statistics received a code level less than 3, and 80% of those who do not teach statistics received a code level less than 3. This indicates that in general instructors utilized very few components in their descriptions and that statistics teaching experience does not seem to produce a different type of response. Similarly, looking at Table 5.8 it becomes evident that the percentage of overall response continues to be consistent even when the instructors' characteristics are considered.

Table 5.8

*Instructors' Justifications That Included a Particular Component In B.Q5*

| Component | Mentioned | All n=12 | M n=5 | MT/ST n=7 |
|---|---|---|---|---|
| A- Context | Yes | 17% | 20% | 14% |
| | No | 83% | 80% | 86% |
| B- Zero frequency | Yes | 17% | 20% | 14% |
| | No | 83% | 80% | 86% |
| C- Data summaries | Yes | 25% | 20% | 29% |
| | No | 75% | 80% | 71% |
| D- Overall trend | Yes | 58% | 60% | 57% |
| | No | 42% | 40% | 53% |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics. A) The consideration for context; particularly the explicit identification of the variable of interest "years in town", B) The ability to clearly recognize that the missing values on the scale corresponded to the zero frequencies on the data set, C) The ability to recognize the limitations of graphical display that does not include zero frequencies, but still recognize that some data summaries could be extracted from such display, D) The ability to recognize that a better scale permits describing the overall trend of a distribution, and E) The ability to take context into consideration when making conclusions (not included in the table because none of the instructors mentioned this component).

Overall the results indicate that instructors tended to ignore, for the most part, the context of the data. Also instructors did not seem to recognize that the uneven scales represent the zero frequency. For example, making reference to Graph 1 of

question B.5, one instructor stated, "messed up graph, they don't know what they are doing" (MT18). These kinds of statements suggest some lack of awareness of students' common mistakes in graphical representations, namely not including non-occurring values (zero frequencies) in the scale (Konold & Higgins, 2002). It also seems that some instructors cannot find any use if the scale is not correct; however, the mode and the range of the data does not depend on the scale. On the other hand, a few instructors gave very thorough descriptions using context primarily to guide their decisions. They also recognized the limitations of Graph 1 but were able to describe the mode and the range of the data without discarding it completely because of the uneven scale.

**Analysis of Question B.3**

Question B.3 (Figure 5.4) aimed to identify if instructors could recognize the best display of the proportion of hits from baseball players so that the shape, the center and the spread of the data could be easily described. The results indicate that many instructors did not recognize the appropriate graph, and even when they did identify the right graph, their justifications revealed that their choice was not based on profound understanding. Moreover, context was for the most part not considered, leading many to the wrong conclusion.

The data were coded first according to the graph chosen. The results indicate that 25% (n=3) of the instructors chose B because of the shape, 17% (n=2) of instructors chose D because it went from the lowest to the highest, and 50% (n=6) of

the instructors chose C. While choosing the right graph was very important, more valuable insight came from instructors' justifications.

**INSTRUMENT B**
**THINK ALOUD**

**Please solve the following problems and while you do so, try to say everything that goes through your mind.**

**3]** A baseball fan likes to keep track of statistics for the local high school baseball team. One of the statistics she recorded is the proportion of hits obtained by each player based on the number of times at bat as shown in the table below. Which of the following graphs gives the best display of the distribution of proportion of hits in that it allows the baseball fan to describe the shape, center and spread of the variable, proportion of hits?

| Player | Proportion of hits |
|---|---|
| BH | 0.305 |
| HA | 0.229 |
| JS | 0.281 |
| TC | 0.097 |
| MM | 0.167 |
| GV | 0.333 |
| RC | 0.085 |

| Player | Proportion of hits |
|---|---|
| SU | 0.270 |
| DH | 0.136 |
| TO | 0.218 |
| RL | 0.267 |
| JB | 0.270 |
| WG | 0.054 |
| MH | 0.108 |

| Player | Proportion of hits |
|---|---|
| BC | 0.301 |
| AA | 0.143 |
| HK | 0.341 |
| RS | 0.261 |
| CR | 0.115 |
| MD | 0.125 |

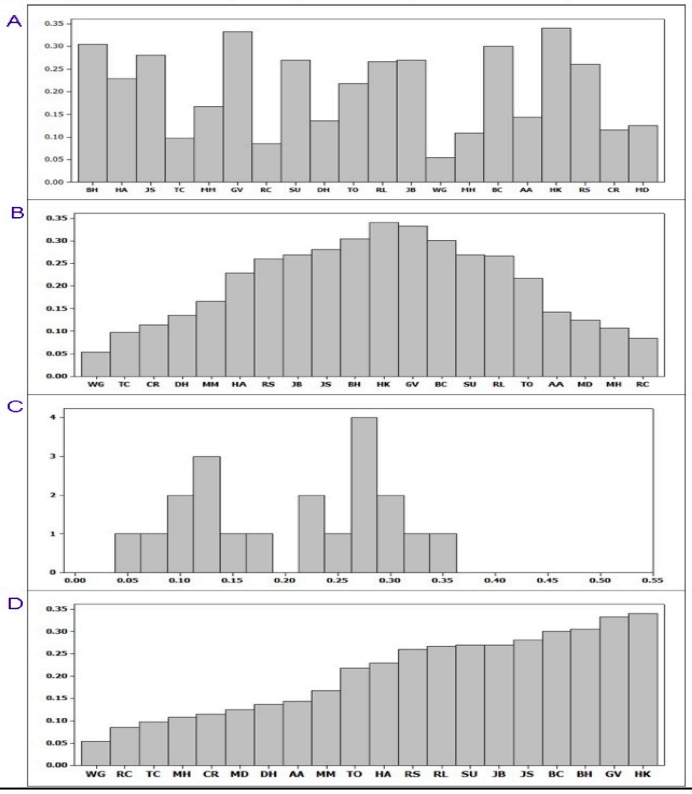



*Figure 5.4.* Instrument B Question 3. Adapted from "The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project" by J. Garfield, R. delMas, and B. Chance, 2002. NSF CCLI grant ASA-0206571. [Retrieved from https://app.gen.umn.edu/artist/].

In order to codify instructors' statements, the researcher first identified the different aspects that needed to be taken into consideration when answering this question. The first was context, the second was to understand the goal of the question itself and then realize that a histogram would be the appropriate display for such data and goal.

Within the context it was very important to understand that the proportion of hits was the data of interest and to therefore recognize that it was a quantitative variable and as such needed to be represented with a histogram. The goal of the problem was to choose a display that would be best to describe the shape, center and spread of the proportion of hits. Whether instructors looked at the goal of the problem (describe, shape, center and spread) or looked at the variable of interest, which is the proportion of hits, the histogram was the only suitable representation, which was choice C in this problem.

This question presented a challenge for instructors. It is possible that the difficulty derived from the fact that several of the graphs presented looked like histograms (A, B and D), as they had bars with something in the x-axis and something in the y-axis, but since the x-axis is not a quantitative measure in any of the three graphs (A,B and D), they are not histograms. Instructors should have recognized that histogram C was the only graph that would address the goal of the problem as well as be the only way to represent the data at hand.

The results were coded as correct if they identified graph C and incorrect if they identified any of the other graphs. The justifications were assigned a value from

zero to four for each of the responses depending on whether instructors mentioned the goal of the problem, that the variable was quantitative, or recognized that C was the only histogram. A code 0 was given to instructors who did not mention any of the three aspects, for example, ST12 only looked at the best way to arrange the data without considering the context, the variable or the goal of the problem.

*ST12:I think this one is the best [pointing at B]. It sort of ranks them from umm well… no actually I take that back… I don't like that one… I change my mind. Well, I… I like it better ranking from lowest to highest… I think that's interesting, although I like that one too… they're both interesting. But I'd say I'd pick probably D. And why? Yeah, yeah am I supposed to say why? Okay… alright… I think it's the… it just it just puts them in order from lowest to highest so I think so one way to represent it. So it's a more organized way. Here they're kind of although… I find that, I find that one interesting but that one's useful too, so… I like this one too [B]… alright let's go to the next page.*

A code of 1 was given if the instructor mentioned only one of the three; for example, MT18 recognized that the data is quantitative and therefore graphs A and B were not of use, but then got confused and did not answer the question at hand.

*MT18; Ok, A and B are terrible because they're using the names of the players and that's not a mathematical scale, so you can reorder them, it's like qualitative data basically that you shouldn't really examine that stuff…I'm confused as to what these even mean, is that the point? I was like, well maybe they're averaging this, but if they're averaging, they'd come up with one number, so…that is what I was going to say, so yes, but, ya, actually none of these are of any use. Ok.*

A code of 2 was given to instructors who mentioned two of these characteristics; for example, MT13 recognized that the variable was the proportion of hits and that only the histogram (C) would be the best way to represent the data, but did not acknowledge the goal of the problem at hand, namely to describe the shape, the center and the spread. It may be understood from the choice of the histogram, but it was not stated and therefore was coded as 2.

*MT13: "The only one that really, if the question is to display the distribution a proportion of hits, then the only one that would make any sense to me is C – that's the only one that even has proportion in the variable"*

A code of 3 was given to instructors who mentioned all three characteristics, for example, MT10 recognized the variable of interest, recognized that the only histogram was C and therefore the only possible graph, recognized the question of interest and even used the histogram to describe the distribution of proportion of hits even though that was not required. She/he seemed familiar with the context and brought that information into the description.

*MT10: So these are numerical values and there is only one [pointing at C] of this graphs that have numerical x-axis is the distribution of proportion is not a distribution of the players so see. So there are two kind of players those into the .05 to and those of the .33 range there are pitchers and batters you know pitchers usually do not there is a lot of pitchers though I can't tell from the initial what position they play so you can tell the shape is bimodal the center is probably like .20 and the spread you can take the standard deviation of that right.*

From the analysis of the 12 instructors' justifications, it is discovered that 50% (n=6) of the instructors received a code zero, while 25% (n=3) received a code three, and 8% (n=1) received code of two and 17% (n=2) a code of one. This means that even within instructors who choose graph C as the appropriate graph, their justifications did not address the key aspects that this question demanded. Of the six instructors who chose C, only three instructors received a code of three while the other three received codes of zero, one, and two respectively. This means that even when the appropriate graph was considered the reasoning does not support a full comprehension of the data at hand.

A possible explanation for the low success rate of this question may be due to the format of the question itself. Out of the 12 instructors who took the interview, 58% (n=7) found the question difficult to understand. Some of the statements are presented below

> *ST3: This is hard to read so this... I am going to assume this is one long table because these are all guys on the team so each person there is a proportion....*
> *MT10: Why there are three tables and four graphs?*
> *MT18: I am confused of what this really means, is that the point?*
> *M16: I can't read this (pointing to the list of data) oh... is this the players' names, player initials (M16 keeps reading) 'that allows the baseball fan to describe?' I don't know what you mean by that, aloud the baseball player to describe? You mean that they can change the way it describes the variable or they can... it gives them a better view of the ordering of what you mean?*

While there may be many reasons why they found the format of the question difficult, the important part to highlight about this difficulty is that this same question with the same format was asked of 1,470 college-level statistics students across the United States with a success rate of 29%. Therefore, the difficulties with the question format expressed by the instructors in this study may help to explain such a low student success rate in this question.

When the results were parsed by instructors' statistics teaching experience it was discovered that there was not a great difference in the type of responses instructors provided. Looking at Table 5.9 it is interesting to note that a large percent (57% n=6) of instructors choosing the incorrect graph had experience in teaching statistics. While it is not clear why this happened, the presentation of question B.3 may need to be investigated. Also, while the percentage of instructors who received a Code 3 in their justification is similar between the groups it is interesting to note

that 72% of instructors who teach statistics received a code less than 3 in their justifications. It may be argued that the characteristics sought in the justifications to this problem were not clearly expressed in the wording of the problem. However, the characteristics sought in their responses are necessary when making decisions with data according to previous research (Konold & Higgins, 2002; Peck, 2005; Shaughnessy et al. 2004).

Table 5.9

*Instructors' Responses to Graph Chosen in B.Q3*

| | M<br>n=5 | MT/ST<br>n=7 |
|---|---|---|
| Graph A | 0% | 0% |
| Graph B | 20% | 43% |
| Graph C | 60% | 43% |
| Graph D | 20% | 14% |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

**Analysis of Question B.4**

Question B.4 (Figure 5.5) was designed to investigate instructors' reactions to students' misconceptions with histograms. In order to understand students' misconceptions, instructors had to first demonstrate consideration for context and realize that the histogram was portraying, in the x-axis, the female literacy rate of South American countries, and realize that the y-axis showed the number of countries with a particular range of literacy rate.

The results indicate that, for the most part, instructors were able to recognize students' errors and correctly identify the reason for students' misconceptions. There are, however, a few instructors who had difficulty with this question.

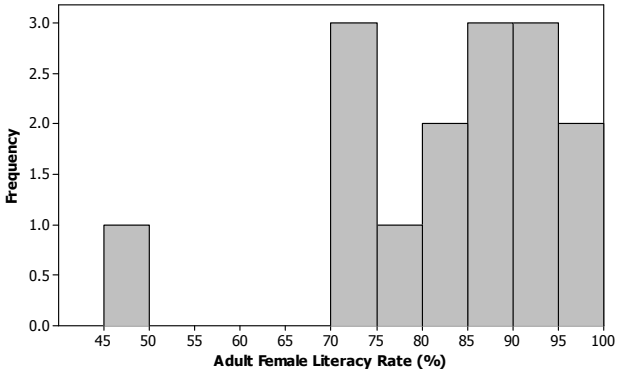
Instructors' responses were first coded for part 'a'; a response was considered correct if instructors recognized that the student was wrong, and also identified the student misconception, namely that the student was counting the numbers of bars instead of adding the frequencies to find out how many countries were represented in the data. Instructors' responses were coded as incorrect if they did not recognize that the student was wrong, or recognized that the student was wrong but could not ascertain the source of the problem. For part 'b', instructors' responses were coded as correct if they recognized that the student gave a partially correct statement but the statement was lacking the frequency. The student should have stated that there were three countries that had between 85% to 90% female literacy rate.

The results of part 'a' indicate that 67% (n=8) of the instructors recognized that the student was wrong and gave the correct reason, 33% (n=4) stated that they could not tell if the student was wrong, or stated that the student was wrong but did not recognize the source of student error. For part 'b', 75% (n=9) of instructors recognized that the student gave only a partially correct description of the histogram and were able to correctly complete the student's description. The other 25% (n=3) of the instructors either incorrectly completed the student's description or did not consider that the student's description was incomplete.

**INSTRUMENT B**
**THINK ALOUD**

**Please solve the following problems and while you do so, try to say everything that goes through your mind.**

**4]** The following graph gives information about the adult female literacy rates in Central and South American countries.




Adult Female Literacy Rates in Central and South America

a. Suppose you ask your students to tell you how many countries are represented in the graph? One student says, “there are 7 countries represented”.

Is this student right or wrong? _____

In your opinion, what is the student’s thinking to arrive to that conclusion?

b. Suppose now you ask your students to explain what the third bar from the right indicates.
One says, “It indicates 85% to 90% literacy rate”.

Comment on the response.

*Figure 5.5*. Instrument B Question 4. Adapted from "Prospective middle school teachers' knowledge about data analysis and its application to teaching" by M. A. Sorto, 2004, Michigan State University, East Lansing.

Instructors' justifications were divided into two parts; those who clearly recognized students' difficulty with the histogram and those who exposed their own misconceptions. Within those who recognized student difficulty with the histograms were clear justifications like:

*MT18 (part a): The student is wrong because the only way that this graph make sense, really, is that the frequency of numbers of countries that have this literacy rate so... The student is wrong. Umm... I would assume that they are counting the number of columns, to come up with that conclusion... Oh! These are... they are thinking this axis [pointing to x-axis] represents the countries, whereas this axis [pointing at y-axis] actually represents the numbers of the countries.*

*ST12 (part b): "it indicates that they were... that they forgot to say that there were three countries in that interval".*

On the other hand, instructors who did not recognize student difficulty with the histograms gave justifications that exposed their own limited understanding of histograms. For example, ST12 stated:

*M14 (part a):"y-axis is telling me the frequency [pointing to x-axis] what is this telling me? This is the age? No, 45 to 50... Make sense the age. Okay, adult literacy rate between 45 and 50. These are the ages. No, this is literacy rate. Don't worry about age. So they are two percent literate age. Well this cannot be the age because there is no info about age. Adult so can be... This frequency... How frequent, like a probability. This is percentage. The student is wrong, in Latin and South America there are more than seven countries"*

M14 (part b): "It is wrong because the high literacy rate in any country should have the lowest frequency".

Overall the results of this question show that most of the instructors (n=8) recognized students' mistakes and also understood the students' thinking. However, it is important to point out that of the four instructors who had difficulty with this

question, two teach statistics (ST17 and MT10). Their full explanations are presented below.

*MT10: You know I always like to shade it, I tell my students this is the only point you can be creative how you shade what colors to chose what patterns you can make don't leave it un-shaded that is what I tell them [reading].*
*So I see they add the frequency 1 plus 3 plus 1 plus 2 that is probably more than seven ah? wrong ah..... students.... 70 in the middle and they move the decimal. I don't know why can have 70 countries but that is my best of what student were thinking. [reading part b']*
*Yes, I think that is right....Oh it is Central and South America. I don't think you have 70 countries there I bet there is only 20 countries total so good intuition student have.*

*ST17: I will say wrong. Student is thinking he or she just added the frequency [pointing to the y-axis and counting]. No does it not it.... [Reading the question]... Conclusion, umm.... I don't know why somebody would have chosen seven, I can't tell why. [Reading part b] is not quite correct, the answer, no very specific, it indicates that the adult female literacy rate in central and South America [...] literacy rate falls between 85 to 90% and the question is not specified the population, for example the female, the female literacy rate in central and South America.*

It is not clear why this difficulty manifested in this question, particularly when it is discovered that MT10 and ST17 were two of the three instructors who received the highest score (three) in the previous question (B.3), which also deals with histograms.

**Summary of Results for Research Question Two**

This chapter was dedicated to answering the question: What statistical concepts do instructors utilize to describe and/or decide the best way to represent one data set and how does context guide those decisions?

Overall it can be deduced that the majority of instructors were able to recognize the appropriate measure of center in the presence of unusual variation (A.Q9). The majority of instructors also recognized the uneven scales in A.Q5 and

B.Q5, but almost none of the instructors indicated that those scales represented only occurring values and that zero frequencies were not included in the display. Moreover, instructors seemed to have a tendency of preferring individual data display instead of display of data as an aggregate (A.Q5).

While several instructors recognized the appropriate display of quantitative data, many struggled to understand the question and/or choose the appropriate graph (B.Q3). It is also evident that the majority of instructors recognized students' misconceptions and were able to appropriately fix the error. However, several instructors showed their own misconceptions when fixing students errors (B.Q4).

The results also indicate that in all the questions presented here the context was the most overlooked characteristic when describing the data and/or making decisions. This in turn affected the type of responses given and therefore several instructors had difficulty with the unusual variation that was presented in A.Q9. Also, lack of consideration for context manifested in A.Q5 where most instructors considered great variability in the display of the graphs dealing with distance in inches representing the braking power of a car, when in reality this was minimal. Similarly, not being able to recognize the variable of interest in question B.3, B.4 and B.5 made several instructors choose the wrong graph and/or not understand what students' problems were.

When the results were organized by instructors' education and statistics teaching experience, there was not any salient difference between the groups. In

general, it seems that the percentage of the results for all the instructors remained consistent even when the results were broken down by instructors' characteristics.

It is interesting to point out that the interview was designed so that instructors could express all of their thinking process without being limited by the space provided to write a response as it was in the survey questions. It was thus hoped that the interview would reveal more depth into instructors' responses. However, the results presented in this chapter reveal that similar patterns of responses permeate all of the questions regardless of whether the question was asked in the survey or the interview.

# Chapter VI

## Results Pertaining to Research Question Three

The analysis of the results presented here aims to answer Research Question Three, along with its sub-questions: What aspects of the graphical representations do instructors pay attention to in the presence of two data sets:

- To decide which graph has more variability?
- To describe and compare the graphical representations?
- To aid them in making informal inference or decisions?
- How are teaching experience and education related to instructors' responses to graphical representations of two data sets?

In order to respond to these questions, this section will describe instructors' responses to four questions of Instrument A (A.4, A.6, A.7 and A.10) designed to reveal their conceptions of variation when dealing with graphical representation of two data sets. There were no interview questions on this topic. The findings presented here are organized in three parts. The first part is dedicated to marking out the features instructors utilized when deciding which graph had more variability. This section will report the results of the analysis of A.Q4 and A.Q10. The second section focuses on reporting what aspects of the distribution instructors pay attention to when dealing with informal inference, as addressed by A.Q6. The third part

illustrates the concepts instructors used when describing and comparing two data sets presented in the form of histograms, which relates to A.Q7.

**Part One: Identifying Graphs with Most or Least Variability**

Questions A.4 and A.10 prompted the instructors to determine which histograms had more variability, and to explain their reasoning. Questions A.4 and A.10a are very similar in the sense that both show two histograms and asked instructors to decide which of the two had more variability. A.Q10b is slightly different since it asked instructors to inspect three histograms and identify which of the three histograms had the most variability and which one had the least variability. The responses for A.Q4 will first be presented, and then for A.Q10. The section ends with the overall results of both of these questions.

**Analysis of question A.4.**

Question A.4 prompted instructors to determine which histogram had more variability and to explain their reasoning (Figure 6.1). The results show that even when giving the correct response, instructors did not always use all aspects of the distributions to substantiate those answers; the majority used only one aspect of a distribution, namely the range. The responses to the first part of the question were coded as being correct if instructors identified School A as having more variability and incorrect if they identified School B as having more variability. The results from this part of the question A.4 seem very encouraging since 77% (n=40) of the

instructors were able to correctly identify that School A had greater variability, while 19% (n=10) incorrectly identified School B as having more variability or decided that both were the same. Four percent (n=2) identified both schools as having the same variability.

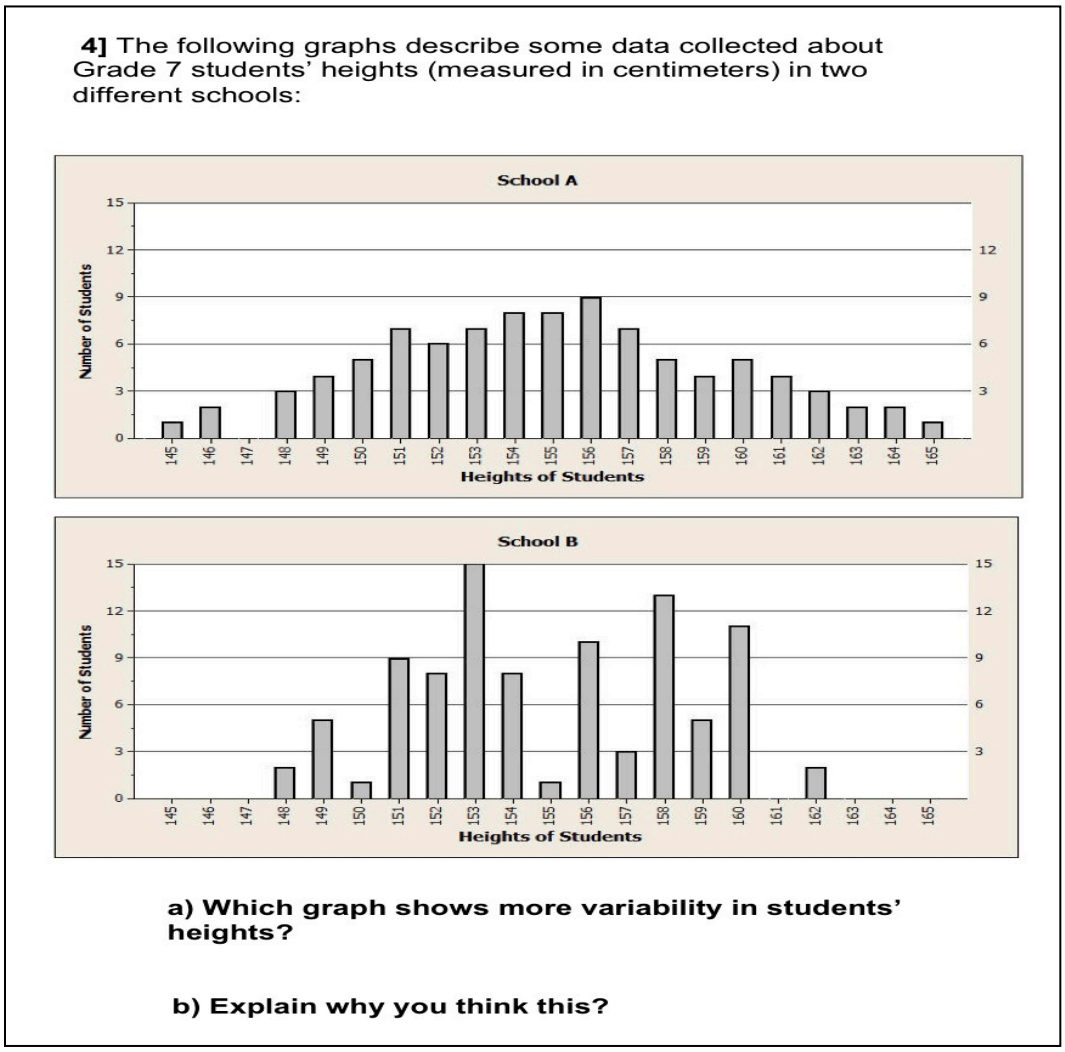



*Figure 6.1.* Instrument A Question 4. Adapted from "Elementary pre-service teachers' conceptions of variation" by D. Canada, 2004, Portland State University, Portland.

While many instructors chose the correct graph, their justifications unveiled limitations and misunderstandings. Table 6.1 presents some of the instructors' chosen graph with their corresponding reasoning.

Table 6.1

*Examples of Instructors' Chosen Graph and Their Reasoning in A.Q4*

| Chosen Graph | Instructors' Reasons | ID |
|---|---|---|
| Graph A | On average the data from school A seems to be farther from the mean than the data in school B. In particular the data at the end of the [..] distribution will make large contributions to the variance. | MT9 |
| Graph A | Greater Range / Though comparing through other measures might give different answers. | ST2 |
| Graph A | Despite School B's student's heights not following a rather Normal Distribution, the range of the scores is lesser, and I would imagine that the means would be similar for both, but the standard deviation would lesser for School B. | MT38 |
| Graph A | The range is greater in "A." I could be wrong since I didn't calculate the standard deviation. | M42 |
| Graph A | Graph A has slight increases or decreases going from each height to the next. Graph B has large jumps going from each height to the next. | M47 |
| Both Same | The first group is more normal, but the second group has a smaller range. | MT38 |
| Graph B | Pure bell vs. "bi-modal" | MT29 |
| Graph B | Because School A is more "normal" in distribution, with less variation. | MT51 |
| Graph B | Groups of data is further apart from the mean in School B | M32 |
| Graph B | School B does not show a normal distribution. There are spikes all over the graph. So even though the range for School B's heights is smaller (148-162) vs. (145-165), I would bet that School B has a greater variance of heights. | M23 |

*Note.* Graph A had greater variability.

From instructors' justifications it was discovered that collectively instructors mentioned several aspects of the distribution: range or spread, extreme values, shape, and spread from center. However, individually these ideas were mostly used in isolation. For example, of those who identified School A as having more variability, 3% (n=1) did not provide any reason for choosing School A, 50% (n=20) of instructors focused on the x-axis or range, ignoring other aspects of the distribution to make their decision, 8% (n=3) of instructors focused only on the shape of the distribution, 6% (n=2) of the instructors showed some misconception, and 6% (n=2) were not interpretable.

Only 30% (n=12) of instructors used a combination of different aspects of a distribution. They either used spread from center, or range and shape, or range and standard deviation. This appeared in that order of frequency. This means that 70% (n=28) of instructors who chose the appropriate graph as having more variability had justifications that lacked consideration of several aspects of the distribution, indicating weak conception of variation when dealing with graphical representations. While the majority of instructors mentioned only the range, which is only partially correct, focusing solely on the range for a measure of variability ignores the influence that the frequencies of each value has when considering variation, which could lead to other problems.

The fact that instructors incorrectly chose School B as having more variability suggests that they had some difficulty with graphical representation and

the concept of variation. To understand some of these difficulties, the researcher looked at their justifications and found some aspects of the distributions they tended to focus on that led them to the wrong decision. For example, 40% (n=4) of instructors decided that school B had more variability because it was not normally distributed. This means that they based their judgment solely on the shape of the distribution, ignoring the range, the spread from center and the frequency of each value. This led them to the wrong conclusion.

Some of the instructors' reasoning was more difficult to comprehend. For example, 20% (n=2) of instructors wrote that their decision was based on the spread of the data from the center, which is the right conception of variation, but they used this understanding to choose the wrong graph (School B). For example, ST22 chose School B as having more variability, and wrote, "top graph shows more concentration around the mean." This instructor seems to ignore the frequency of the values because there are more observations in School B in the center range (150 to 160) than in School A and there are more observations outside that range in School A than in School B. His/her judgment is therefore unsubstantiated.

When the data is parsed by instructors' backgrounds (Table 6.2 and Table 6.3) it is found that the percentages of instructors selecting the appropriate graph with more variability are similar, with a small difference shown in instructors who have never taught statistics (M). Overall, question A.4 unveiled some interesting aspects of the conception of variation of two-year college instructors when

identifying graphs with greater variability. First, the identification of the appropriate graph with more variability does not guarantee the appropriate reasoning.

Table 6.2

*Instructors' Answers to Which Graph Shows More Variability?*

| | All<br>n=52 | M<br>n=23 | MT<br>n=22 | ST<br>n=7 |
|---|---|---|---|---|
| Same | 2% | 0% | 2% | 0% |
| School B | 19% | 30% | 4% | 14% |
| School A | 79% | 70% | 86% | 86% |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics.

Table 6.3

*Justifications for Choosing the Graph With More Variability A.Q4*

| | | MT/ST<br>n=29 | M<br>n=23 | All<br>n=52 |
|---|---|---|---|---|
| School A | Range | 45% | 40% | 42% |
| | Shape | 0% | 9% | 4% |
| | Combine | 28% | 17% | 22% |
| | Other | 14% | 4% | 10% |
| School B | Y-axis | 0% | 13% | 6% |
| | Normal | 7% | 4% | 6% |
| | Other | 3% | 13% | 8% |
| Both Schools Same | | 3% | 0% | 2% |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics.

Second, while some instructors seemed to grasp the idea of variability as a measure of spread about the center, a few still used this knowledge inappropriately by identifying the wrong graph. Third, it seems that there is an incorrect association of seeing the normality shape of a distribution as an indication of low variability.

**Analysis of question A.10.**

Question A.10 (Figure 6.2) was also designed to assess instructors' ability to identify histograms, which had greater or smaller variability, and to find out what aspects of the distribution they paid attention to for such identification. This question had two parts: part a asked them to identify which of two histograms had greater variability, and part b asked them to identify which of three histograms had the most variability and which had the least variability.

As in question A.4, several instructors recognized Class G in question A.10a as having greater variability, but their justifications focused mostly on the range of the distribution and showed a limited consideration for variation. The answer was coded as being correct if the instructor identified Class G as having greater variability and incorrect if they identified Class F as having greater variability or stated that both had the same variability. About 48% (n=25) of instructors were able to correctly identify Class G as having greater variability, while 44% (n=23) said that Class F had greater variability, 2% (n=1) said that they were equal, and 6% (n=3) did not respond. It is interesting to note that even though the question is similar to Question A.4, a smaller percentage of instructors identified the correct

graph in this question. This suggests that the limitations found in instructors' previous justifications are now becoming apparent in their inability to recognize the graph with more variability.

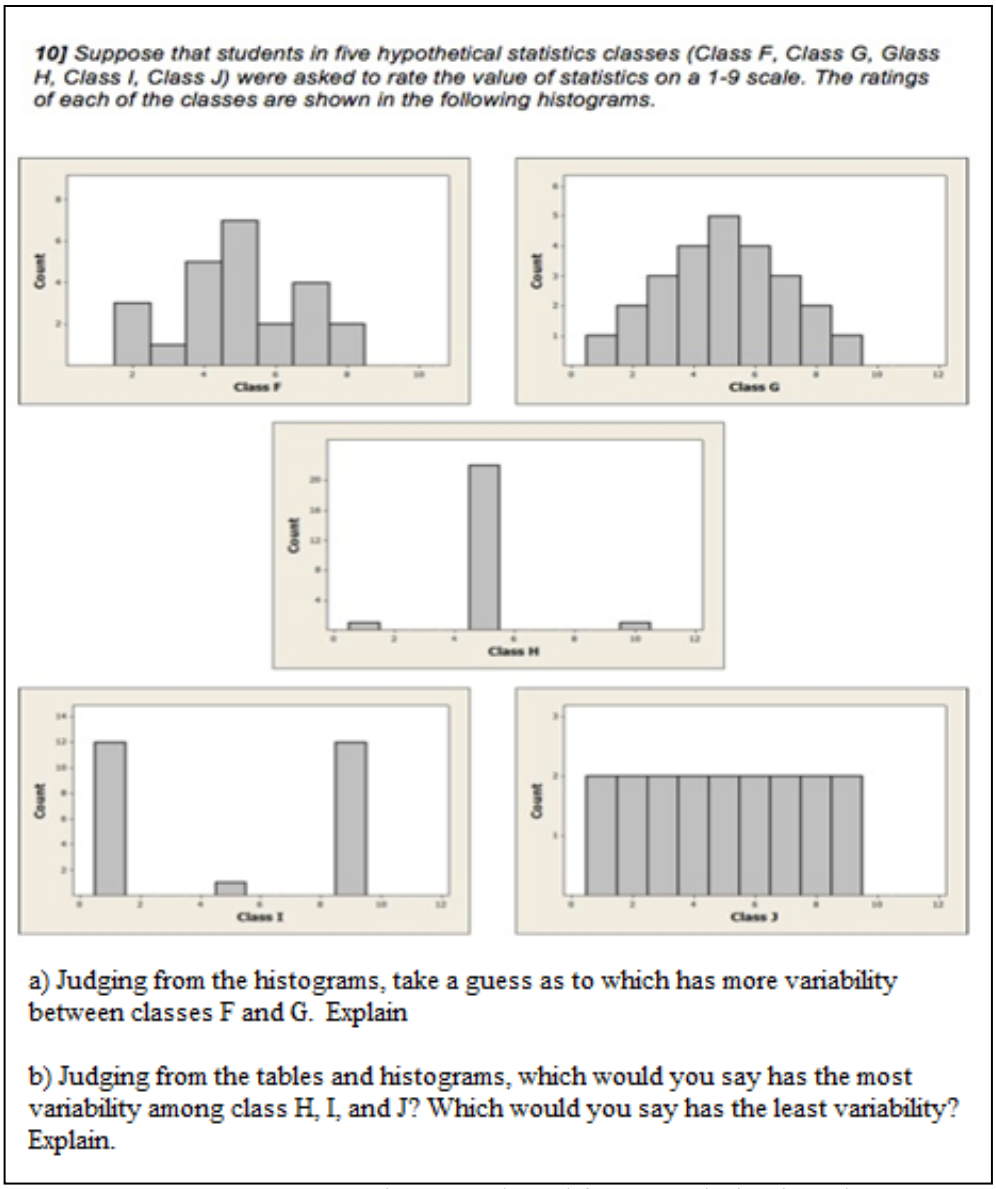



*Figure 6.2.* Instrument A Question 10. Adapted from "Exploring introductory statistics students' understanding of variation in histograms" by M. Meletiou-Mavrotheris and C. Lee, 2005, in paper presented at the Fourth Congress of ERME, the European Society for Research in Mathematics Education, Sant Feliu de Guíxols, Spain.

The justifications given to this question followed a similar pattern as in the previous question in the sense that instructors continued focusing on the x-axis or range as the sole determinant for recognizing variability, or used other aspects of the distribution in isolation. For example, looking at the justifications given by instructors who correctly identified Class G as having more variability, 52% (n=13) of instructors focused only on the x-axis or range of the distribution. Only 24% (n=6) of instructors used several aspects of the distribution. For example, some used spread from center or shape and center, and some even mentioned the range and the spread from center of the data. Of the 25 instructors who correctly identified Class G, 20% (n=5) gave no reasoning or stated that they were guessing and/or that they did not know and 4% (n=1) focused only on the shape of the distribution.

These percentages indicate that while the majority was able to recognize the appropriate graph, they gave very little consideration for all aspects of the distribution. It may be argued that range was sufficient evidence to determine which one had greater variability because Class G had larger range than Class F, and therefore instructors did not have the need to mention other aspects of the distribution. Yet the lack of consistency in the answers seems to indicate that this limited view (range only) can indeed lead them to the wrong decision when identifying the graph with more variability.

Instructors who indicated erroneously that Class F had more variability included justifications that showed misconceptions. About 43% (n=10) can be described as not mentioning any aspects of the distribution, while 57% (n=13) of

instructors who described aspects of the distribution showed misconceptions. Within the reasons that did not include any aspects of the distribution, it was found that 13% (n=3) of instructors stated that they guessed, 9% (n=2) were not interpretable, and a large percentage (22%, n=5) did not justify their choice, thus making it difficult to understand why they chose the wrong graph.

Within instructors who mentioned some aspects of the distribution, it is found that 30% (n=7) of the instructors indicated Class F as having more variability because the tails had more values. For example, MT24 wrote, "F because there are more values out to each extreme in the x distribution"; however, instructors who gave this kind of response did not recognize that the tails of Class F are closer to the center, making the variability smaller even if there are more values in them. Also, about 9% (n=2) indicated that Class F had more variability because it was not normally distributed. Another 9% (n=2) said, "F has more variability because there are larger jumps"(M47), meaning that they are looking at the y-axis as a measure of variability, and 9% (n=2) used the idea of spread from center but used it to identify the wrong graph.

Overall, it can be deduced that those who identified the graph correctly tended to focus solely on the range of the distribution and those who identified the graph incorrectly tended to focus on the frequency of the extreme values. Paying attention to the range and to the frequency of the extreme values are both good considerations to determine the amount of variability in a graph. Using them in isolation, however, can lead to wrong conclusions.

Table 6.4

*Percent of Instructors Who Identified Graph with Most Variability in A.Q10a*

| | M | MT | ST | All |
|---|---|---|---|---|
| | n=23 | n=22 | n=7 | n=52 |
| Class F | 39% | 45% | 43% | 42% |
| Class G | 43% | 55% | 43% | 48% |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics. Class G had greater variability. The total does not add up to 100% because one instructor stated that both were equal, and four did not respond.

Looking at Table 6.4 it is clear that the percentage of instructors who correctly identified the graph with the most variability does not vary much when comparing instructors' education and statistics teaching experience.

While question A.10a gave instructors two histograms to look at and to decide which graph had greater variability and to justify the response, question A.10b presented three different histograms and asked instructors to state which one had the most variability, which one had the least variability and to justify that decision. This question had, in essence, three parts that were embedded in the same sentence, which may explain why 90% (n=47) of the instructors answered the first part (which graph had more variability), and only 56% (n=29) answered the second part of the question (which graph had the least variability); this means that 44% (n=23) ignored or chose not to respond to the second part of A.Q10b.

The results show that for the most part instructors had no difficulty identifying Class I as having the most variability and Class H as having the least variability. Class J seemed to pose a challenge since many instructors incorrectly identified Class J as having the most or least variability. However, within the graphs presented in this question, Class J had neither the most nor the least variability. As in the previous cases, the justifications confirmed that instructors tended to focus on a single aspect of the distribution, making it difficult for them to assess or justify the variability of Class J.

The response was coded as being correct if they stated that Class I had the most variability and Class H had the least variability. Those who stated it backwards or chose Class J as either having the least or most variability were coded as being incorrect. The results show that the majority of the instructors (63%, n=33) were able to recognize that Class I represented the class with the most variability, while 25% (n=13) incorrectly identified Class J as having the most variability. Six percent (n=3) said Class H had the most variability and the remaining 6% (n=3) did not respond.

For the second part of this question (which class had the least variability), it was found that 48% (n=25) of instructors were able to recognize that Class H had the least variability, while 6% said that Class J had the least amount of variability and 46% (n=24) did not respond. It can be inferred from these results that the histograms presented by Class I and Class H did not seem to be a challenge since a great percentage identified them correctly.

On the other hand, Class J, which had the histogram with the uniform distribution, was erroneously identified by 31% (n=16) of the instructors; they either said that Class J had the least amount of variability (n=4), or as the majority (n=12) believed that Class J had the most variability. It seems that the idea of looking at only the range and ignoring the spread from the center of the distribution as a measure of variability led these instructors to the wrong decision. Those who identified Class J as having the least variability were probably focusing on the y-axis.

The justifications given by the instructors revealed that instructors who focused on describing how the data is spread from the center tended to correctly identify both the graph with the least variability and the graph with the most variability. Instructors who paid attention only to the x-axis or the y-axis tended to identify the graphs incorrectly. Overall, 35% (n=18) identified both Class I as having the most variability and Class H as having the least variability, while the remaining 65% identified the graphs incorrectly or did not respond to both parts of the question.

Table 6.5 shows the percentage of instructors who recognized the graph with the most variability. When looking more closely at instructors' education and statistics teaching experience it becomes very evident that there seems to be a difference in the way they respond. It appears that there is a relationship between their background and the type of response instructors give, as confirmed by the test statistic result, $\chi^2(2, N = 52) = 10.56, p = .005$. A similar pattern emerges with instructors who identified the graph with the least variability, but the percentages are

not presented since a large part of the group did not respond to this part of the question.

Table 6.5.

*Percent of Instructors who Identified Graph with Most Variability in A.Q10b*

| | M<br>n=23 | MT<br>n=22 | ST<br>n=7 | All<br>n=52 |
|---|---|---|---|---|
| Class I | 39% | 82% | 86% | 62% |
| Other than Class I | 61% | 18% | 14% | 38% |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics. Class I had the most variability.

The interesting part of this finding is that as long as the graphical representations could be judged solely on the x-axis or range there seemed not to be a distinction in the responses according to instructors' education background and statistics teaching experience; this was shown in A.Q4 and in A.Q10a. However, when the distributions can not be judged by the range, it seems that a natural separation occurs between those who grasp the idea of variability and those who do not, and it seems that those who have experience teaching statistics have developed a more sound understanding of the idea of variability as a measure of the spread from the center.

It is important to note that there were some unexpected justifications from instructors who teach statistics, particularly when considering the influence that these instructors have on the students they teach. For example, ST17 wrote "J, b/c has no variability". From his/her statement it is not clear what part of the question was being answered, but it appears that he/she was looking at the y-axis to justify the

lack of variability that he/she is expressing. Also, MT5 wrote "Class J the most more # chosen & class H the least b/c most people answered 5". It appears that this instructor was looking at the x-axis and the frequency in which the numbers appeared in each of the cells, but disregarding the idea of the spread from the center, which led him/her to the wrong decision.

**Similarities between questions A.4 and A.10.**

A.Q4 and A.Q10 are similar in that they both prompted instructors to determine which histogram had more variability and to explain their reasoning. Some instructors were consistent throughout the items in providing answers that reflected an appropriate idea of variability, namely considering spread from center; however, the majority showed inconsistencies in their responses. Comparing the responses from A.Q4 and A.Q10a (Table 6.6), it becomes evident that instructors were not consistent in identifying the graphs with greater variability. It is not clear why a large percent of instructors were able to identify the variability in some graphs and not in others.

Table 6.6.

*Percent of Instructors Who Correctly Identified the Graph With More Variability*

| | A.Q4 n=52 | A.Q10a n=52 | A.Q10b n=52 |
|---|---|---|---|
| Identified correctly | 77% | 48% | 62% |
| Identified Incorrectly | 23% | 46% | 29% |
| Did not respond | 0% | 6% | 9% |

When looking at the justifications, some patterns of responses became evident. Instructors who correctly identified the graph with more variability gave responses that can be divided into those who focused on only one aspect of the distribution, mostly the spread, and those who used a combination of aspects of the distribution. Those who focused on a single aspect tended to focus on range. While one way to look at variability is through the range, paying attention to only that aspect of a distribution can prove limiting and may lead to the wrong conclusion, as found in A.Q10b. Those who used a combination of aspects tended to pay attention to the spread from the center, which proved in most cases to be sufficient to correctly identify the graph.

On the other hand, different kinds of responses were discovered from instructors who incorrectly identified Graph B as having greater variability in A.Q4 and Graph F as having more variability in A.Q10a. For example, three main categories emerged: instructors who paid attention to the frequency of extreme values, or looked at the y-axis; instructors who judged variability based on the shape of a normal distribution; and some instructors who used the idea of spread from center to incorrectly identify the graph with more variability.

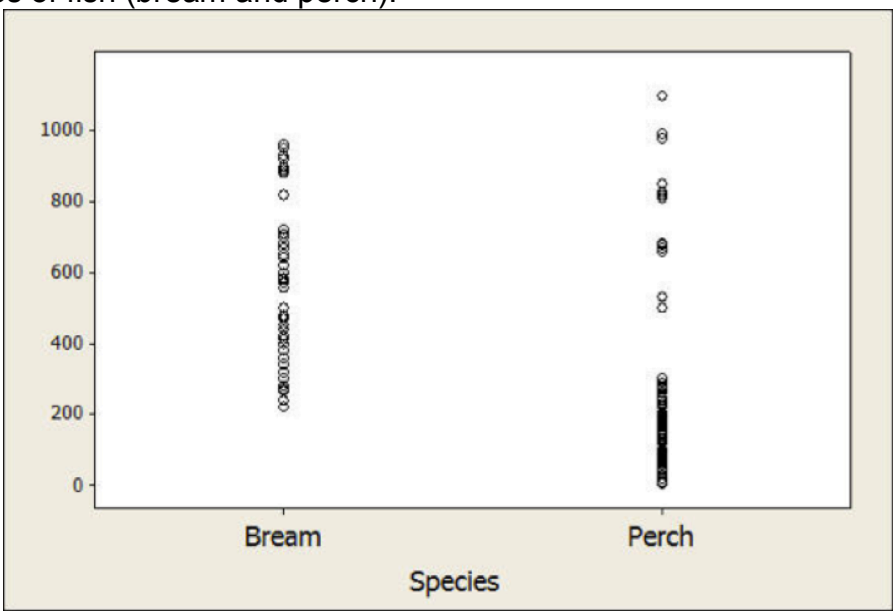
## Part Two: Informal Inference

### Analysis of question A.6

Another graphical representation of two data sets was presented in question A.6 (Figure 6.3) with the purpose of gaining insight into instructors' reasoning about

variation in a more sophisticated context. Here instructors not only needed to pay attention to the variability found in one data set (each type of fish), but they needed to juxtapose the variability within (spread about the center and frequency of the values) with the variability between the two groups (the center of the data) to reach some informal inference about the weight difference of the two fish. In other words, they needed to decide whether there is a meaningful difference in the weight relative to the distribution of the weights.

**6]** Look at the following plot. It shows the weights in grams of two species of fish (bream and perch).




a) Do you think that there is a difference in weight for the two species? Explain your response.

*Figure 6.3.* Instrument A Question 6. Adapted from "Reasoning about variation: Student Voice" by C. Reading & J. Reid, 2007, *International Electronic Journal of Mathematics Education*, 2(3), pp. 110-127.

The results indicate that there were two main types of responses: those who made some decision (either that there was a difference or that there was not), and those who did not make a clear decision about the weight of the two species. The data was coded first for whether the instructor either explicitly or implicitly said something about weight difference. About 75% (n=39) of the instructors did recognize a difference between the two species, 8% (n=4) stated that the two species were the same, 15% (n=8) of instructors expressed doubt without making a decision, stating that "it depends", and 2% (n=1) did not respond.

A thorough analysis of instructors' justifications revealed that instructors who expressed doubts (15%, n=8) seemed confused about how to look at this question and stated that it "depends on whether you mean a difference in the mean weights of the species, or a difference in the variation of weights of the species" (ST48), or "maybe, not averages, but yes standard deviation. The perch have larger range" (ST3). It seems that they could not integrate all aspects of the distribution to make their decision.

Of the instructors who made a decision, a clear division emanates between those who stated that there was not a difference, and those who stated that there was a difference. Instructors who stated that there was not a difference (8%, n=4) mentioned only the average and did not utilize any descriptions of spread or variability to justify their decisions; for example, ST17 wrote, "On average, they seem to have the same weight" and MT7 wrote, "Averages appear to be equal".

These descriptions are also vague since the instructors did not clarify what they meant by “average”.

The justifications of instructors who stated that there was a weight difference (75%, n=39), on the other hand, fell into five categories: no justification; justifications that did not mention any aspect of the distributions; justifications that focused on different aspects of distributions in isolation without specifying anything about weight difference; justifications that concentrated on one aspect of the distribution; and justifications that can be identified as separating the signal (the bulk of the data) from the noise (the variation around the signal) (Konold or Pollatsek, 2004) and using the signals to make their decision.

Of the instructors who stated that there was a difference in the weight of the two species, 21% (n=8) did not specify what aspects of the distribution they were looking at and simply wrote, "seems like overall the Perch is lighter" (MT28), "yes the perch weigh less" (M34), or "looks like the perch is lighter" (MT29). About 10% (n=4) described characteristics for each distribution and made no further connection to what they decided about the weight difference of the two species of fish; for example, they stated " Perch: greater variability at the lower values/ Bream more symmetric" (ST1), or "yes. Bream tends to have evenly distributed weight, while Perch has more data stacked near the bottom of the data range" (M32). It seems that they were unable to make a decision about whether there is a difference in the weight of the fish, but instead looked at the difference in the distributions without stating what they thought about the weight. Of the instructors who stated that there was a

difference in the weight of the two species, 26% (n=10) focused only on the density of the distribution. They mentioned how the concentration of the majority of the perch was at the lower weight values. For example, they wrote, "yes, because most of the weights of the perch are below 400 grams" (MT31), or "yes there seem to be more perch at lower weight than bream," (MT6) indicating a focus on only the density of the distribution.

Of the instructors who stated that there was a difference in the weight of the species, 26% (n=10) were able to separate the signal from the noise. The ability to separate the signal from the noise derives from looking at several aspects of the distribution to determine what the signal is on each, and then comparing the signals without being distracted by the noise. For example, MT51 wrote, "yes, the perch is definitely lighter, since the clear majority are in the 50-300 g range. The bream looks pretty evenly spread out between 200-900. The perch has a few outliers, which wouldn't raise the average much". This shows that these instructors were able to consider both the variation within each of the distributions and the variation between the two distributions, and to then judge that the variation between had precedence even in the presence of extreme values.

Overall, even though a large percentage (75%) of the instructors recognized a difference between the weight of the two species, only 40% (n=21) of instructors mentioned some aspect of the distribution in their justifications to aid them in their assessment of the fish weight difference. This means that a majority of the

instructors demonstrated, at best, a very weak conception of variation when it comes to making decisions based solely on graphical representations.

Looking at the characteristics of these 40% (n=21) of instructors who were able to express aspects of the distribution to aid them in their decision, it is discovered that 62% (n=13) have a degree in mathematics and have never taught statistics, while the other 38% (n=8) have both a degree in mathematics and experience teaching statistics. Since none of the seven instructors who have a degree in statistics were among those who clearly expressed what they thought about the difference of weight between the two species, their responses are presented in Table 6.7.

Several of these instructors mentioned variability and even standard deviation, but for some reason they seemed to find a conflict between looking at center versus spread. However, as was explained earlier, the goal was to assess the characteristics of the graphical representations and decide how the variation within each data set may affect the variation between the data sets so that some kind of informal inference about the weight difference could be reached. It seems that these instructors (ST) are separating the variation in between (as they mentioned the mean), and the variation within (as they mentioned standard deviation), but they are not integrating them to reach a conclusion; some of the instructors expressed that doing a formal statistics test would be the path to take.

Table 6.7

*Justifications to A.Q6 from all Instructors With a Statistics Degree*

| Instructor Justification | ID |
|---|---|
| Perch: greater variability denser at the lower values / Bream more symmetric | ST1 |
| Clearly there are differences in the samples for one, the bream sample is less variable | ST2 |
| Maybe not averages but yes standard deviation. The perch have a larger range | ST3 |
| Bream has a higher mean/median weight/ Perch has [..] a higher standard deviation | ST12 |
| I would need to run a test to see if the difference is significant (with specified level of confidence) But If I were to guess, I would say that there is no difference. On average, they seem to have the same weight. | ST17 |
| Yes - overall it looks like perch is lighter, if looking at percentiles, not averages | ST22 |
| Depends on whether you mean a difference in the mean weights of the species, or a difference in the variation of weights of the species. | ST48 |

The goal of this question (A.6) was to answer what aspects of the distributions instructors pay attention to in order to aid them in making decisions (informal inference) based only on the graphical representations of two data sets. The majority of instructors reached some kind of inference; whether they found a difference or not, they did come to some kind of decision about the weight difference between the two fish. However, the results indicate that instructors utilized few

aspects of the distributions to aid their decisions about weight difference of the two fish. Those who did pay attention were able, in some cases, to differentiate the signal from the noise, while others paid attention only to the shape or density of the distribution. Those who focused on the average stated that the weights were equal for both fish, leaving little room to understand what aspects of the distribution they were focusing on.

## Part Three: Describing and Comparing Distributions

### Analysis of Question A.7

Question A.7 (Figure 6.4) was used to find out what aspects of a distribution instructors pay attention to when describing and comparing two graphical representations. The written responses were first scanned to identify any explicit language for describing a distribution, particularly dealing with measure of center (mean, median and mode) and spread (range, spread, shape, IQR, standard deviation or variance). The result of this indicates that only 27% of the instructors used some of these terms, while 73% did not use any of these terms explicitly. The most common expressions used to compare the two bus routes was by describing that one was more reliable than the other, or one was more consistent than the other, or one had greater variability than the other.

Given that the language that the instructors used was not explicit in describing aspects of center or spread of the distribution, the data were coded again, as follows: Firstly, identifying any measure of center that instructors had mentioned

in their written responses (explicitly or implicitly), then secondly identifying any measure of variability (explicit or implicit) and finally identifying the use of both of these measures together.

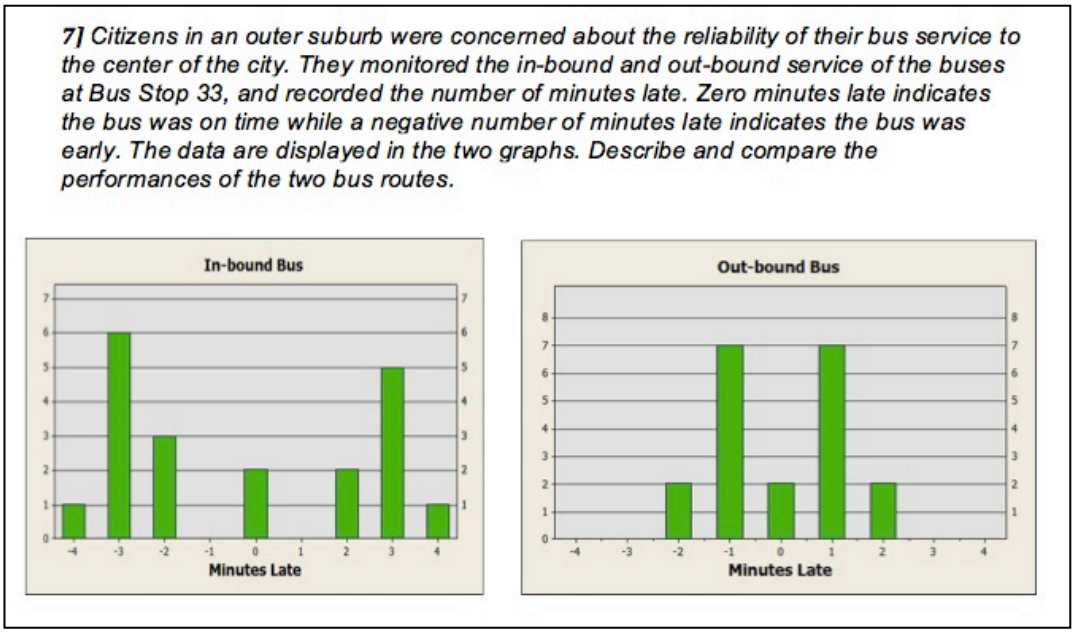



*Figure 6.4.* Instrument A Question 7. Adapted from "Reasoning about variation: Student Voice" by C. Reading & J. Reid, 2007, *International Electronic Journal of Mathematics Education*, 2(3), pp. 110-127.

Two different forms of addressing the center were discovered in instructors' responses: one was coded “explicit” and the other was coded “implicit”. Within the explicit measures of center were words like the mean, the median, the mode and average. Examples of explicit mention of center includes statements like, "their mean were both 0 (right on time)" (MT9), or "both mean= median=0" (ST1), or "both sets seem to have average 0" (MT6).

Implicit measures of center included references to being on time, early, or late. These statements were coded as implicit consideration of center because in the context of the data given being on time was zero and it was at the center of the distribution. An example of implicit consideration of center includes statements like "the out bound more on time" (M36), or "usually within +/- 1 minute of its schedule time" (MT39).

Measures of spread were coded as explicit variability, implicit variability, or vague variability. Within explicit variability were words like range or spread, for example, "Both seem fairly reliable although in-bound buses have a fairly large spread from being 4 minutes late to 4 minutes early" (M19), or "out is more reliable in that the results show less variability (measuring variability through the range)"(ST2).

Within implicit variability were statements like "the inbound was early or late by up to 4 minutes. The outbound bus was early or late by up to 2 minutes" (M41), which implied the concept of spread. Also within implicit variability were statements like "the out-bound bus seemed to adhere to the schedule more closely" (M42), which implied that they were considering the distance of values from center, or "in bound bus was early 10 times, on time 2 times, and late 8 times. The out bound bus was early 9 times on time 2 times and late 9 times" (M43), which implied that they were taking the frequencies of the values near the center or away from the center into consideration.

Within vague variability were statements like "in-bound times vary much more than the out-bound times" (MT8), or "outbound- more consistent- less variability" (ST1). It is clear that they were considering the variability of the distribution, but it is not clear what aspects of the distribution they were paying attention to and therefore it was coded as vague variability.

The results show that 15% (n=8) of instructors explicitly mentioned the idea of center (mean, median, or mode), while 56% (n=29) implicitly referred to the center (on time, early, or late) and 23% (n=12) did not mention the center either explicitly or implicitly and 6% (n=3) did not respond at all. This means that the majority 71% (n=37) of instructors made some reference to the center of the distribution in their descriptions, but they were for the most part making implicit rather than explicit reference to the center.

With regard to variation, the results show that the majority 87% (n=45) of the instructors described some form of variability in their responses. However, 38% (n=20) of all instructors were coded as vague variability, while only 25% (n=13) made explicit reference: four mentioned the range, four mentioned the spread, two mentioned the variance, one mentioned the standard deviation, one mentioned the frequency and one mentioned the deviation from the center. Of all the instructors, 23% (n=12) made implicit reference to variability, while only 8% (n=4) did not mention any form of variability and 6% (n=3) did not respond at all.

Describing some form of center and spread are essential in describing any distribution, and more important is the ability to integrate both ideas of center and

spread to compare the distributions (Canada, 2007). However, several instructors used both of these ideas but did not connect them, while others were able to integrate them. Of all the instructors, 50% (n=26) used them independently, as shown by statements like "the means of the two sets are the same, but the inbound bus times appear to be more variable" (MT21); this shows that the instructor mentioned the mean and also wrote about the variability but did not integrate them.

Of all the instructors, 40% (n=21) were able to combine center and variability, as shown by statements like "Same number of bus trips. Outbound bus deviates less in their arrival times compared to the schedule arrival times. Thus it is more reliable. The in-bound bus arrival times fluctuate a great deal more and the early arrivals and late arrivals can be 2 times those of the out-bound bus." (MT5). Here it becomes evident that the instructor referred to the spread around the center of both of the distributions to make his/her judgment. Note that about 10% (n=5) of the instructors did not mention either center or variability.

Another aspect that it is interesting to mention about instructors' responses is that 21% (n=11) of all instructors mentioned something about the context connecting it to something personal; for example, M16 wrote:

*In-bound: Early a lot but also late a lot. I would like taking this bus. Since they can't leave before the schedule time anyway, being early is of little value. Out-bound: very consistent. I would like this bus. No more than 2 minutes late in 20 trials, and mostly within one minute of time, sounds good to me.*

While the 11 instructors who mentioned context in their statements made different kinds of connections with the idea of riding a bus, it is interesting to note

that those who mentioned the context more readily expressed ideas of center and spread of the distribution. For example, eight out of the 11 who mentioned context integrated both center and variability, as in the case of M16 mentioned above.

When looking at the results, it may be tempting to suggest the lower percentages are due to the fact that many instructors do not have a background in statistics and do not teach statistics. However, when the data were parceled into instructors' characteristics (Table 6.8), it becomes evident that each group has about the same percentage of instructors using implicit or explicit measures of center and of variability in the three groups (M, MT and ST) and this follows the same pattern of response as the entire group. This means that instructors' education and statistics teaching experience do not seem to be related to the way instructors answer these questions.

Table 6.8

*Mention of Centers in Justifications for AQ7: Percent by Group*

| | M<br>n=23 | MT<br>n=22 | ST<br>n=7 | Total<br>n=52 |
|---|---|---|---|---|
| Explicit | 13% (3) | 18% (4) | 14% (1) | 15% (8) |
| Implicit | 61% (14) | 55% (12) | 43% (3) | 56% (29) |
| No center | 17% (4) | 23% (5) | 43% (3) | 23% (12) |
| No response | 9% (2) | 4% (1) | 0% | 6% (3) |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

Table 6.9

*Mention of Variability in Justifications for AQ7: Percent by Group*

| | M<br>n=23 | MT<br>n=22 | ST<br>n=7 | Total<br>n=52 |
|---|---|---|---|---|
| Explicit | 22% (5) | 27% (6) | 29% (2) | 25% (13) |
| Implicit | 30% (7) | 18% (4) | 14% (1) | 23% (12) |
| Vague | 26% (6) | 45%(10) | 57% (4) | 38% (20) |
| No variability | 13% (3) | 5% (1) | 0% | 8% (4) |
| No response | 9% (2) | 5% (1) | 0% | 6% (3) |

*Note.* M = math degree, does not teach statistics; MT = math degree, teaches statistics; ST = statistics degree, teaches statistics

Overall, question A.7 revealed that instructors did not use very explicit language to describe the distributions and compare them but rather used implicit language to describe the center and the variability of the data. A small percentage of instructors were able to integrate the ideas of center and spread, a theme that seems recurrent in these types of questions. While very few instructors (n=11) mentioned the context, it seems that they tended to integrate the ideas of center and spread more readily. Another recurring result in this study is that the differences found in the type of responses do not seem to be related to instructors' education or statistics teaching experience.

**Summary of Results for Chapter 6**

The main purpose of this chapter was to present the results of questions designed to answer Research Question Three: What aspects of the graphical representations do instructors pay attention to in the presence of two data sets,

- To decide which graph has more variability?
- To describe and compare the graphical representations?
- To aid them in making informal inference or decisions?
- How are teaching experience and education related to instructors' responses to graphical representations of two data sets?

The results indicate that the responses given by instructors were categorized in two forms: the short answer responses and the justifications. Throughout the questions presented in this section (A.4, A.10, A.6 and A.7), the short answer responses seemed to indicate that instructors have good consideration for variation. However, delving more deeply into instructors' justifications seemed to reveal an array of limitations. When looking at the data in this way it becomes evident that throughout these questions a small percentage of instructors seemed to utilize several aspects of a distribution to aid them in their decision and/or descriptions of the data at hand.

The aspects of the distribution that were the most commonly used overall was the concept of spread (minimum value to the maximum value) and the idea of range (maximum value minus the minimum value). While these two considerations

of variation are important, when they were used in isolation they led some instructors to choose the wrong graphs. This became most evident in A.Q10b where the instructors were presented with a histogram that had a uniform distribution. Recall that 31% (n=16) of the instructors identified this uniform distribution incorrectly. The justifications shed light on the fact that instructors in some way or another associated a normal distribution with less variability compared to one that is not normally distributed.

When they were asked to assess if there was a difference in the weight of two species of fish (A.Q6), instructors seemed to have difficulty in discerning how to answer the question. Several instructors gave vague explanations or were not able to integrate spread and center. It seemed that the unfamiliarity with informal inference may have confused them, with some expressing the need to perform a formal statistical test.

When it came to describing and comparing graphs, instructors expressed ideas of center and variability more implicitly than explicitly. Many utilized aspects of range or mean, but in isolation. Integrating the center and the spread as a measure of variability was evident in some cases. It is interesting to note that questions in which the graphical displays bear similarities tended to produce a similar kind of response. For example, each of these questions presented in this section (A.4, A.6, A.7 and A.10) required similar thought processes and ideas of variability. One aspect that becomes evident is that when there is an obvious mark of the center of the data (as shown in Class H, Class I in A.Q10, and also in A.Q7), instructors are more

ready to use the idea of spread from center; however, when the evidence of the center is not clear (as shown in Class J in A.Q10, and also in A.Q4 and A.Q6), instructors do not seem to apply the idea of spread from center.

The results also reveal that for the most part there seems not to be a difference in the responses according to instructors' education and statistics teaching experience; of all the questions analyzed here, only A.10b shows a significant difference between the groups.

# Chapter VII

# Discussion

Much information was gained about instructors' conceptions of variation in this study. This chapter is dedicated to discussing some of the findings including inferences drawn from the results, implications and limitations, which will be addressed in turn for each research question. The chapter finishes with general implications, general limitations and comments about future directions.

## Discussion Pertaining to Research Question One

### Variability in repeated samples.

The results of the questions dealing with repeated samples (A.1, A.2, A.3, B.1 and B.2) indicate that not all instructors answered the questions in the same way. Some expressed consideration for variability in the numerical responses when taking samples and corroborated their understanding by giving clear justifications about their consideration of variability. Others, however, did not. The difference is not explained by statistics teaching experience or even by college degree held.

After reading the results of chapter four dealing with repeated samples, it may be concluded that there is nothing that anyone can do right in statistics when it comes to repeated samples. Almost every possible answer appeared to have a

potential fault. For example, answers that were 666666 were criticized for lack of variability, answers with 123456 were criticized for being too low, 6678910 for being too high, 556677 for being too normal, and answers with 0 to 10 were criticized because of the somewhat implausible vision that anything is possible.

It seems that there are few plausible answers. However, therein lies the core understanding of variability. What can make one suspect that something is completely off or not? One may wonder what the fine line is between theory and practice, and who has the appropriate understanding to use such a precise measuring tool. Reading and Shaughnessy, (2004), acknowledge that "random variation is puzzling even to statisticians and researchers" (p. 202). Many of the instructors' responses appeared to be an application of statistical theory rather than a consideration of the practical issues that were used to frame each of the questions.

Many instructors in this study used the theoretical formulas formula either explicitly or in their heads. For example, A.Q1 would be $E(x) = 10*0.6 = 6$, a response that was given by a number of people. Many of the instructors' responses to A.Q3 were also based on probability theory. In theory, a particular face of a fair die has a 1/6 chance of occurring. In theory, the probability distribution of the faces of a die follows a uniform distribution. Therefore, when instructors had to predict what would happen in 60 trials, they possibly used the formula for the expected value $E(X)=n*p$, producing $E(X)=60*(1/6)=10$. Since the trials were independent they expected each face to come up 10 times. It is therefore not a surprise that the result

of (10,10,10,10,10,10) was given by the majority of the instructors when answering A.Q3.

Understanding of the interplay between theory and practice becomes even more necessary when dealing with empirical sampling distribution as instructors were faced with in B.Q1. As the results indicate, none of the 12 instructors were able to discern between the graphs and correctly identify all the four graphs as either fake or real. One of the possible reasons for the inability to correctly identify these graphs may be grounded in instructors' knowledge or lack thereof of the theoretical sampling distributions. First, considering the context of B.Q1 one can say that the distribution will follow an approximate (since the original population is finite) binomial distribution. Therefore one would expect (theoretically) 7 times out of the 50 times to get 6 red candies in a handful of ten candies.

Question B.1 was done as an experiment with candies, which adds another layer of variability into the picture. However, it is essential to understand that the empirical distribution (dealing with the candies) is closely related to the theoretical distribution of 50 samples. Sampling distributions is not necessarily an easy topic to master, particularly if one is not trained in statistics. One way to gain confidence in how empirical distributions relate to theoretical distributions is to perform simulations. By doing several simulations the eye gets trained to see how theory and practice relate. Most likely, the instructors that took part in this study have not been exposed to such simulation as only three of the 12 instructors in the interview part of the study had a degree in statistics. It is interesting, however, that none of the

instructors with a degree in statistics were able to identify the four graphs correctly and they did not use all aspects of a distribution to aid them in their judging.

The question that naturally follows is why do instructors have such difficulty with this question? Even though this study is only exploratory in nature it brings to light results that, if they were to hold true for all two-year college statistics instructors, would indicate that there may be a greater problem to solve. Instructors at two-year colleges are required to teach sampling distribution. It is on the basis of the understanding and mastering of this topic and its interconnection to concepts of distribution and center that the concept of statistical inference lies.

It is important to understand, however, that the researcher did not inquire about instructors' thinking or prompt them to clarify what they were verbalizing their responses during the interview. It was a Think-aloud process where instructors explained what came to mind without the researcher's intervention. To get a clearer picture of instructors' thinking, a follow-up interview with interaction between the participants and the researcher may well provide a better view of instructors' conceptions of variation when dealing with repeated samples.

**Connection to previous research.**

The results from this study were very similar to studies done previously. Reading and Shaughnessy (2004) found it intriguing that students who did NAEP (National Assessment of Education Progress) in 1996 provided single-value responses to a sampling task. In that particular assessment, out of 232 students

sampled only one provided a range of values when students were asked how many red gumballs they would expect in a handful of ten. It became evident that students do not volunteer ranges for this kind of question. These researchers found this "troubling because it suggests that students don't recognize the role that variability plays in a sampling task" (p. 208). If Reading and Shaughnessy considered it troubling that 4th grade students gave this kind of response, then how should one consider the fact that 87% of the college-level instructors in this study gave a single value response for the sampling questions?

If one analyzes the responses provided by instructors in this study in the light of previous research that utilized the same items (A.Q1 and A.Q3) (Canada, 2004; Reading & Shaughnessy, 2000, 2004; Watson & Shaughnessy. 2004), one may be tempted to suggest that the main reason for providing little consideration for variation is due to the emphasis on probability theory and the tendency of focusing on the expected value, as several researchers have proclaimed (Canada, 2004; Reading & Shaughnessy, 2000, 2004; Watson & Shaughnessy, 2004). However, even though this point of view may be valid and applicable to the two-year college instructors that took part in this study, it may not be the only explanation as to why there is a tendency to respond with expected values. It is possible to postulate that the questions themselves may have led participants to these kinds of responses.

Reading and Shaughnessy, 2004, provided a perfect example of how the different words used in a question suggest the difference between asking for theory versus the results from an experiment. They explained that:

> The probability question is: What is the range of all possible outcomes and which ones are more likely to occur than others (i.e., what is the sample space?).
> The statistical question is: if we repeat a probability experiment many times, what sort of outcomes do we observe, and what is the range of the more likely outcomes (i.e., what interval captures most of our trials?) (p. 206).

These two questions highlight the difference that would result in different kinds of responses. However, it may be safe to assume that instructors, for the most part, have not been exposed to both sets of questions in a way that highlights the differences between them. On the contrary, probability theory formatted questions inundate not only the probability theory courses to which most instructors are exposed, but also the instructional textbooks used by instructors teaching statistics, which for many are the only reference to statistics knowledge. If one considers that in this study 87% of the instructors do not have a degree in statistics and have in most cases only textbook exposure and previous courses on probability theory as a resource, it comes as no surprise that they seemed not to distinguish the difference between a probability question and a statistics question, and chose to respond with the typical answers that do not acknowledge variability. This does not mean that they do not have developed conceptions of variation, but rather that the distinction between these types of questions has not been highlighted anywhere.

**Implications**

Further research is needed to answer some of the questions that this study has brought to light. It is possible that a more interactive kind of interview would reveal other levels of instructors' thinking. Also, the type of question seems to have an influence on the type of response and therefore further research is needed to include questions that reveal more clearly instructors' understanding of center and the appropriate spread from center. Some questions did not bring out the instructors' conceptions of variability (A.1a and A.3), while in question B.2 instructors clearly expressed most important considerations about a distribution, including the appropriate amount of variability around the center. It seems imperative to investigate further the effect that the questions themselves may have in allowing access to instructors' (or students') knowledge.

The results of this study are also relevant for teaching since it becomes evident that not all questions trigger the same knowledge. For example, asking only short type responses does not seem to reveal the depth of the understanding of a particular concept; it is therefore recommended that justifications accompany short responses when assessing students' thinking. In order to assess variability in students, the type of questions themselves need to be carefully selected since some may be better than others to reveal students' conceptions of variation, as was discovered in the instructors' responses.

The items used in this study may also be utilized as a springboard for discussion in the classroom since they seem to elicit different kind of responses. The

differences in the responses are worth considering as they may help to explain the difference between theory and practice.

## Discussion Pertaining to Research Question Two

### One data set

One of the goals of this study was to gain insight into how instructors dealt with one data set. From the results, it became evident that several instructors recognized the presence of the outlier in A.Q9 and used appropriate measures of central tendency. Several instructors, however, insisted on using the arithmetic mean even in the presence of an outlier. It was discovered that in general instructors were able to recognize students' difficulty with histograms, but several instructors revealed their own misunderstandings with histograms when trying to identify students' errors. Also, several instructors found it difficult to match a quantitative data set (proportion of hits in B.Q3) with the appropriate graphical representation, the histogram.

It was interesting to discover that most instructors have a preference for graphs that show individual data points rather than graphs that summarize the results. Moreover, the results indicate that the majority of instructors did not identify the gaps in the scales of the graphical display as zero frequencies, and that they tended to discard a graphical representation based solely on the uneven scale of the graph.

One of the most overlooked characteristics that instructors exhibited in the analysis and description of one data set was the lack of consideration for context.

This was evident in all of the questions. The lack of consideration of context led many of the instructors to wrong conclusions. Another interesting aspect of the results was that the types of responses instructors provided were similar across instructors regardless of their education background and/or their statistics teaching experience.

These results bring to light some issues that need to be explored. Several explanations could be offered to help understand some of the issues revealed in this study. For example, instructors' idea of the mean being a robust measure, even in the presence of an outlier, may be due to the fact that most formal statistics tests (two sample t-test, ANOVA, etc.) are grounded on the mean. The sample mean is also considered an unbiased estimator for the population mean.

While these conceptions of the mean are correct, they hold true under certain conditions that were not present in the question they were answering (A.9). For example, in order to use the sample mean as representative of the population, the sample has to either be taken from a known normally distributed population, or the sample size has to be large (more than 30). However, in the question presented here the population was unknown and the sample size was only ten. Therefore, while the justifications used by instructors hold true under specific circumstances, they did not in the question at hand. This may indicate a lack of practical experience on the part of the instructors because most researchers would be very concerned with the presence of an outlier in a small data set.

In general, instructors recognized the appropriate graphical representation for the proportion of hits (B.Q3), and were able to identify students' misconceptions when interpreting histograms. However, the fact that several instructors struggled with this graphical representation raises the question of why. One possibility is to consider whether there is a problem with the questions themselves. This idea needs to be particularly entertained as several instructors stated that the question (B.3) was not clear. One possible explanation for this difficulty could be the way the row data was presented. Typical introductory statistics textbooks present raw data in a list format that includes a single quantitative variable, or a single list of qualitative variables, or a list with two quantitative variables, but rarely are lists found with several variables in which some are qualitative and some quantitative. It seems that qualitative and quantitative variables are usually kept separate.

It may be argued that simply changing the way the raw data was presented in question B.3 would have produce different types of results. For example, if the raw data only included the quantitative variable (the proportion of hits) and excluded the player names, it is possible that instructors would not have been confused about the type of graphical representation needed. However, while it is plausible that the way the data were displayed could be a possible source of the problem, it is important to highlight that consideration of the context and the variable of interest should have been sufficient to choose the appropriate graphical representation regardless of the display of the data.

Another issue that came to light is the fact that instructors preferred graphical representations of individual data points rather than displays that represent summary of data. This is a tendency also found in students, and researchers suggest that students do not recognize the importance of seeing the data as aggregate until they need to use graphical representations for comparison between groups (Konold & Higgins, 2002). However, it is likely that instructors have been exposed to such graphical comparisons. Therefore, it can be deduced that the lack of consideration for context may have been the greater obstacle when those decisions were made. The idea that information is lost when data are grouped together, together with the idea that such displays mask variability, seems to be the opposite of what it is expected for students to learn. This type of reasoning may require further investigation if the goal is for students to move from seeing data as individual to seeing data as an aggregate.

This study also found that the majority of the instructors seemed unaware of the reason for the uneven scales in the graphical displays. Some even blamed students for "messing around" or "not knowing what they are doing". However, there is evidence in prior research that students tend to graph only occurring data points, excluding categories on the scale for non-occurring data points (Konold & Higgins, 2002). It seems that instructors had a similar problem as students and could not understand why the scale had the gaps.

It also became evident that uneven scales were considered by instructors to be completely misleading. While this type of conclusion may be valid, it is more

important to ascertain the usability of a display by considering the question of interest that the display will help to answer (Konold & Higgings, 2002). For example, displays with uneven scales can still reveal the mode of the data as well as the range. So if that was all that was being sought to answer then the display of uneven scale would be sufficient.

While questions that would ask only for the mode or the range of the data are rare in college, the fact that instructors did not consider the question of interest to decide that the graphs were misleading seems troublesome. If flexibility with the displayed graphs is expected from elementary and middle school students as Konold and Higgins (2002) have stated, then two-year college instructors should have at least a similar level of awareness and flexibility.

Overall the biggest problem highlighted in the results of this section is the lack of consideration for context. As Roxy Peck clearly states "in statistics, meaning comes from context, and it is the interpretation of the analysis in context that is the ultimate desired outcome of analyzing data" (p. 1). She suggests that statistics courses typically tend to focus on computational aspects of the data without putting too much emphasis on the context. This understanding may be the explanation for why the majority of instructors (even those with a degree in statistics) ignored the context for the most part.

**Connection to previous research.**

The questions discussed in this section deal with one data set in several representations varying from a list of individual data points, a display of individual data points, and displays of data as aggregates as in histograms or dot plots.

Several researchers have utilized question A.9, which was presented as a list of individual data points. The groups of individuals investigated with this question were diverse. Some researchers utilized it to investigate prospective teachers (Sorto, 2004), others utilized it to investigate middle and high school students (Watson et. al, 2003), while others utilized it to investigate college students (Garfield, 2002). Regardless of the population studied in prior research, the studies revealed that students as well as prospective teachers tend not to consider throwing out the outlier and averaging the rest of the numbers, but rather insist on adding all the numbers including the outlier and dividing them by the total. In trying to explain such a phenomenon, Sorto (2004) concluded, from the results of an item similar to question A.9 in this study, that the "the attraction to the mean (Method III) appears stronger and more complicated than first thought" (Sorto, 2004, p. 135). This study revealed that even instructors who teach statistics and have a degree in statistics insisted on using the mean even in the presence of outliers, thus adding evidence to support the pervasiveness of the phenomenon.

Individual data displays appeared in questions A.5 and B.5. Prior research dealing with these items revealed that students tend to describe individual cases in the data instead of using summaries. For example, a student would say, "someone

lived there 37 years"(Watson et al., 2003). It has been recognized that students have a tendency to talk about the numbers without making reference to the context. For example, students would simply say "four x on 3" (Watson et al., p. 16). Another aspect that came to the surface with the individual data displays was that as soon as the data were clustered or grouped, instructors tended to think that it was less informative. However, the benefits of data as an aggregate include the ability to see emergent features, like center, spread and shape. These features are lost when looking only at individual cases.

Several studies have been conducted to assess college students and prospective teachers with similar questions dealing with histograms as presented in this study. For example, prior research showed that students' success rate in question B.3 was very low. Question B.3 was part of the Compressive Assessment of Outcomes for First Course in Statistics (CAOS), which was administered to 1,470 introductory statistics students from 39 different colleges in the United States. The results indicated that only 29% of the students were able to recognize the appropriate graphical representation for the number of hits. While the success rate of instructors for this question was higher at 50%, several instructors failed to recognize the appropriate graph even though some of them teach statistics. Further investigation may be needed to understand whether the format of the question itself may be influencing the results, or the lack of exposure to raw data lists that include mixed variables (qualitative and quantitative), or a combination of both. It is important to highlight in reference to the format of question B.3 that some of the instructors in

this study stated that the question was not clear. They did not understand why there were four lists of data, and they were also confused with having two variables in the list. The types of issues that instructors brought up about the format of question B.3 may explain some of the low success rate that students have experienced in the past.

When dealing with histograms (B.Q4), Sorto investigated prospective teachers and found that the majority (75%) of the prospective teachers were able to recognize students' error, but they gave incorrect ways to correct the mistake. She also found that several prospective teachers did not recognize the meaning of the y-axis in a histogram. She attributed this difficulty to the fact that only the x-axis had a label for the variable, Literacy Rate (%), while the y-axis said only 'frequency'.

Moreover, Meletiou-Mavrotheris and Lee (2005) in their analysis of college students found that 70% of the students gave wrong interpretations of the histogram. They attribute this difficulty to the fact that the raw data were transformed into a new form of representation. This study of instructors has similarity to previous findings, even though the success rate of instructors in this study was higher, but the fact that many of them seem to have difficulty with histograms raises an interesting question: Could the questions themselves be the problem?

**Implications**

The way the instruments were administered could be improved in order to better understand the instructors' reasoning. For example, instead of a survey and an individual interview, instructors could take part in a group interview in which the

questions are solved and discussed in groups. An alternative could be that they solve the questions independently and then meet to discuss their results. This type of interactions would open the door for deeper understanding of instructors' conceptions of variation as their thoughts may be revealed when discussing the different ways to answer a particular question. From the results of this study it can be inferred that not all instructors answered the questions in the same manner, therefore the opportunity to openly discuss different views could bring a deeper understanding into instructors' thinking. This proposed methodology could be a springboard for enriching discoveries into instructors' thinking that could then also be utilized to understand students' thinking.

It is possible that the questions themselves need to be revised and reformatted to assess statistics knowledge. It is important to highlight that to thoroughly assess somebody's knowledge requires more than 16 questions. The results presented here give only a glimpse of the way that instructors think, although multiple questions were used to tap into most of the concepts.

Given the poor showing on many of these questions, it seems reasonable to ask if researchers are using the right questions to assess statistical knowledge? Since the questions for the survey and interview utilized here were all part of previous studies conducted mostly on students, it is important to question the validity of the questions themselves when instructors do not respond as would have been expected. It is likely that instructors know more than these questions seem to reveal. Therefore, prior conclusions drawn about students' statistics knowledge that are based on these

questions need to be seriously reconsidered if the goal is to understand students' thinking about variability.

There are a number of instructional implications from this study. Activities that would reinforce the effect of unusual variability on the measure of center should be implemented. The questions presented here and used in several studies could be used in class to discuss with students the different ways to deal with outliers and to also discuss the benefits of one method over the other.

Since histograms seem to present so many difficulties, it is proposed that histogram displays should include labels clearly stating the variable of interest on the x-axis and also the variable on the y-axis. The y-axis represents the count, or number of occurrences, which is why the y-axis is label 'frequency'; however, if the goal is to imbed statistics in the context of the data at hand, then the y-axis has to explicitly state what is being counted. For example, in the two histograms presented here (B.Q3 and B.Q4) the y-axis could be labeled 'Number of Players' (for B.Q3), and 'Number of Countries' (for B.Q4). This would avoid confusion and the context would be accentuated. It seems easy to fix and it could clarify not only the meaning of the frequency but also bring the context to life

Another way to improve understanding of histograms may be to expose students to raw data lists that include several mixed variables, since in real data collections more than one variable is normally considered. This could also bring awareness that the appropriate display needs to match the type of variable and the question being asked.

## Discussion Pertaining to Research Question Three

### Two data sets.

Overall the results to questions A.4, A.6, A.7 and A.10, dealing with graphical representations of two data sets, indicate that the short answer types of responses tend to produce results that obscure instructors' conceptions of variation. The justifications were more revealing about instructors' conceptions of variation since they had to explain the reasoning behind their choices.

The aspects of the distribution that were the most commonly used in the different scenarios were the concept of spread (minimum value to the maximum value) and the idea of range (maximum value minus the minimum value). Moreover, when instructors were asked to assess if there was a difference in the weight of two species of fish (A.Q6) several instructors gave vague explanations or were not able to integrate spread and center.

When it came to describing and comparing graphs, instructors expressed ideas of center and variability more implicitly than explicitly. Many utilized aspects of range or mean, but in isolation from each other. Integration of the center and the spread as a measure of variability was evident only in some cases.

The fact instructors' justifications in many cases did not support the short answer responses could be due to the fact that choosing responses generally requires less thinking than explaining responses. Therefore, by asking for justifications of

their choices this study has reached a deeper level of understanding of instructors' knowledge of variation.

Another discovery in this study was that many instructors paid attention to the range to ascertain which graph had more variability. This may be partially due to the fact the range was sufficient evidence for such a decision, and the range is in fact a measure of variability. However, it could also be because instructors only equate variability with range. What was noticed in the results is that instructors' showed less difficulty distinguishing which graph had more (or less) variability when the distributions shared some salient feature.

For example, when the range was obviously different between the distributions (A.Q4), 70% of the instructors identified the graph correctly. When the spread from center was obvious (A.Q10b), 63% of the instructors identified the graph correctly. However, when the difference in range or the difference of spread from center was not very pronounced (A.Q10a), then the correct identification of the graph with the most variability (or least) decreased to 48%. This means that when the difference in range or the difference in the spread from center is well-delineated, instructors seem to experience little difficulty. However, distributions with more subtle differences produced greater difficulties.

Questions that required comparing distributions proved to be the most challenging for instructors. Instructors tended to base their decisions on a single aspect of a distribution or to provide vague statements that hid their conceptions of variation. Specifically, it seems that they were not able, for the most part, to express

variability as the spread from center and were unable to separate the signal and noise to extract the essence of the data. It was discussed in the previous section that instructors dealing with a single data set demonstrated several difficulties. It is understandable that they would experience greater difficulties when needing a more sophisticated approach to compare distributions and performing some kind of informal inference.

One possible explanation for the difficulties could be attributed to the fact that introductory statistics courses traditionally focus on formal inference, which tends to be limited to computational procedures and 'cookie-cutter' responses (Peck, 2005, p. 2). It is then not surprising that instructors struggled to make some kind of inference without the accustomed statistical numerical summaries like the mean, the standard deviation, or the sample size which are needed for formal statistics tests (two sample z-test, two sample t-test).

Researchers are increasingly emphasizing the need to focus less on computations and more on exploring and discerning what is happening with the data. Peck (2005), referring to students, stated, "if all a student can bring to the table is the ability to perform computational aspects of data analysis, they can be easily replaced by an inexpensive calculator" (p. 1). It seems that there is a call from researchers to move away from simple computation to a more sophisticated understanding of the data at hand.

**Connection with previous research.**

Prior studies have concluded that it is difficult to help students develop the concept of variability as spread from center (Garfield, 2007). In this study the dependency on the range as a measure of variability became evident. Prior studies (Reading & Reid, 2006) that investigated students' conceptions of variation have developed hierarchies of students' reasoning. While the hierarchies have many levels, students who mentioned only the range as a characteristic of variability were considered as being at the lowest level of such hierarchies, namely a "weak consideration of variation" (p. 48). In prior research (Garfield, 2007) it was found that the idea of distribution needed to be a strong foundation before students were able to appropriately deal with measure of variability. The responses presented by instructors in this study seemed not to reflect such an understanding of distribution. There appears to be a gap between what students need and what instructors have mastered.

Question A.10 of this study was adapted from a prior study of college students conducted by Meletiou-Mavrotheris and Lee (2005). They asked students to identify graphs with more variability but their study was conducted in several stages. They first asked which graph had more variability before teaching a class on histograms and variability (Pre-instruction). They then taught a class and asked the same question with a different graph that showed similar distributions as before the lesson (Post-lesson). They concluded by using a computer program to find the summary of statistics for the same graphs that were presented to them in the Post-

lesson stage (Post-Computer). What is interesting to note between these two studies (theirs and this one) is that the percentages of success most similar to this study are the percentages of students' responses from the Pre-lesson. This possibly indicates a natural disposition about the ideas of variability. It is proposed that the idea that seems to be appearing throughout this study is that some students, and some instructors (regardless of their characteristic (i.e. M, MT, or ST), seem to easily grasp the idea of variability while others need further instructional aids to achieve such awareness.

Table 7.1

*Comparison Of Results Of Prior Study With This Study*

| Identification of graph with more variability | Meletiou-Mavrotheris and Lee study (Students' responses) | | | This study (Instructors' responses) |
|---|---|---|---|---|
| | Pre-lesson | Post-lesson | Post-computer | |
| Incorrect | 45% (n=15) | 70% (n=21) | 13% (n=4) | 44% (n=23) |
| Correct | 40% (n=13) | 30% (n=9) | 87% (n=26) | 48% (n=25) |

*Note.* The results are based on the short answer responses to the graphs of question A.10a that were used in both studies.

Meletiou-Mavrotheris and Lee (2005) did not discuss students' justifications in detail but in their study one of the justifications for incorrectly choosing the graph

with more variability was that students focused only on the shape of the distribution, believing that a normal distribution had more concentration around the center and therefore less variability. This is exactly the same justification found in this study. Also, students who correctly chose the graph with more variability tended to focus only on the range of the distribution, which was common in this study as well. Meletiou-Mavrotheris and Lee (2005) concluded in their study that the most salient characteristic of their results was that students focused on the y-axis as a measure of variability. Interestingly, only a few instructors focused on the y-axis in this study. However, even though few instructors focused on the y-axis as a measure of variability, some of the instructors did and that may further jeopardize students' understanding of histograms, where it seems that students need the most help.

**Implications.**

The results of this and prior studies seem to suggest that there is a split between those who grasp the idea of variability in graphical representations and those who do not. It may be interesting to further investigate all types of people without a background in statistics to see if this holds true in general. Understanding prior awareness of ideas of variability may help not only with teaching, but also enhance further research in the pursuit of understanding students' and instructors' conceptions of variability.

Comparing graphical representations that do not have salient features seems to be more telling about the depth of conceptions of variability, especially when a

justification is also given. Developing such graphical representations may help unveil deeper levels of understanding of the conceptions of variation and help discover what other aspects instructors (and students) tend to focus on in the absence of salient features. It was discovered in this study that justifications to the questions were key in understanding the depth of instructors' thinking. Instructors should include not only questions that require a short answer type of response, but a section with the justifications where the real level of understanding can be better assessed.

Graphs with more variability could be built by utilizing the questions presented here in a progressive manner. For example, an instructor could first introduce a graph with an obvious difference in the range (A.Q4), then a graph with an obvious difference in the spread from center (A.Q10b), and finally many graphs where the difference is more subtle so that in-depth considerations of variability can be achieved.

**General Implications of Findings**

One of the serious implications from the findings of this study is that the state of statistics knowledge of mathematics instructors at two-year colleges may be far from the knowledge required for effective teaching. According to Shulman (1986), instructors need to have not only content knowledge but also the knowledge necessary for teaching. He stated "the teacher need[s] not only understand that something is so; the teacher must further understand why it is so." (p. 9). Having content knowledge is considered an essential ingredient for teaching. However, the

goal for those teaching is to have not only content knowledge but also the deeper conceptual knowledge necessary for teaching.

However, as presented in this study, the instructors' incomplete and in several cases inaccurate responses are indications that these instructors do not seem to possess the knowledge necessary for teaching. It may be argued that this study included instructors who do not teach statistics and that therefore they should not be expected to have the necessary knowledge for teaching for statistics. However, almost all of the results of this study indicated that the difficulties and limitations of the explanations were found across the different groups (M, MT, and ST). It may also be argued that surveys tended to be done rather quickly and therefore cannot be taken as authoritative. However, the follow-up interviews in this study, where instructors had no time constraints, showed similar results. This means that the results appear to have credibility.

The consequences of instructors' limited conceptions of variability have significant repercussions for students. For example, it is known that students have great difficulty with histograms; however, this study revealed that several of these instructors could not identify students' misconceptions when interpreting histograms. Even though not all of the instructors had prior experience teaching statistics, their limited understanding of simple histograms jeopardized their ability to identify students' misconceptions.

It is important to note, however, that it is not clear who actually has a well-developed sense of variability. A division seems to have occurred in this study where

some instructors readily identified variability and others did not. The difficulties with variability before this study have been identified in several populations. Garfield and Ben-Zvi, (2007), stated, " it is extremely difficult for students to reason about variability" (p. 22). Another group identified as having similar difficulties with variation is K-12 teachers (pre-service and in-service) (Garfield & Ben-Zvi, 2007). Teaching assistants (statistics graduate students) have also shown many similar difficulties to students (Canada 2004). Garfield and Ben-Zvi, (2007), stated that "inappropriate reasoning about statistical ideas is widespread and persistent, similar at all levels even among some experienced researchers, and quite difficult to change" (p. 4). The instructors in this study do not seem to be the exception to the rule. A comprehensive conception of variation appears to be some kind of rare knowledge, even among those who teach at the college level. Is it possible that only a few possess it, or is it the way researchers are trying to assess it that has shown itself to be limited? If only a few possess it, how can the change take place from rare (very few understand it) knowledge to general (everyone at all levels understands it) knowledge through improved teaching?

This study suggests further research could clarify if the questions utilized in this and prior studies are valid tools for measuring variability. If the questions are proven not to be the problem, then deeper investigation into instructors' knowledge of variability may be required. It is suggested that the community of researchers pay greater attention to this population (instructors of two-year colleges) that has been overlooked in past research. This study was exploratory in nature and while there

could be faultfinding in the methods and procedures used, instructors' direct quotes utilized throughout the study should give at least a glimpse of the state of statistical understanding of two-year college mathematics instructors. This study can serve as a springboard for further in-depth studies where the inclusion of interactive interviews, using both clarifying and prompting, would improve the depth of the knowledge exposed in this study. Group interviews could also be beneficial since not everyone seems to hold the same opinion and the interactions and discussion with several instructors with different viewpoints could be very useful for understanding the depth of instructors' thinking.

It is important to recall that this study included only responses from volunteer instructors. However, the result of this study have demonstrated instructors' limitations and should be of concern when considering that more than half of the instructors who took part in this study are currently teaching statistics to thousands of students. Nevertheless, it is important to acknowledge aspects of the study that may account for some of the findings.

First, the wording of the questions and/or the presentation of the data in the questions may have been factors. The questions used in this study were items widely used by other researchers. However, that in itself does not guarantee that the questions were appropriate in measuring what they were meant to measure. Further investigation of the wording and/or presentation of the questions may be desirable.

Another aspect that may have influenced the instructors' responses is the fact that the researcher conducting this study has a statistics background and was a

college mathematics instructor. This may have resulted in some degree of intimidation and/or feeling that for the researcher the answers to the items were maybe obvious and/or easy. Therefore, some of the instructors' uncertainty and lack of confidence could be a result of the researcher's background.

"Important research issues arise with analysis of instructors' justifications, including whether the researcher is interpreting the responses correctly and whether the instructors are responding optimally" (Reading & Reid, 2007, p. 113). To deal with this potential weakness, the researcher embedded as many of the instructors' originally statements as possible in the report. It may be easily assumed that instructors were not responding optimally because there is more depth to their thinking and understanding than a written sentence or a half hour interview could explore. Additional studies are necessary to arrive at a conclusive opinion on the conceptions of variation held by two-year college instructors, but it is hoped that this study has opened the gate for future research into this population.

**Comments about Future Directions**

The findings of this study trigger many possibilities for future research into instructors' reasoning about variation. This study has only scratched the surface. The researcher believes that a written survey and the think-aloud could be accompanied by a personal interactive interview, which may help to further clarify instructors' thinking about variability. Group interviews may also provide an effective tool to further understand their thinking. Whatever path is taken to investigate mathematics

instructors' conception of variation at two-year colleges, it would be significant progress to further understand this overlooked population. Researchers need to seriously consider this population if the aim is to serve students and the community at large.

## References


Aquilonius, C. (2005). *How do college students reason about hypothesis testing in introductory statistics courses?* Unpublished Doctoral Dissertation, University of California Santa Barbara, Santa Barbara.

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

Bataneo, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (2000). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology, 25*(4), 527-547.

Biehler, R. (1997). Students' difficulties in practicing computer-supported data analysis: Some hypothetical generalizations from results of two exploratory studies. In J. Garfield & G. Burrill (Eds.), *Research on the role of technology in teaching and learning statistics* (pp. 169-190). Voorburg, The Netherlands: International Statistical Institute.

Canada, D. (2004). *Elementary preservice teachers' conceptions of variation.* Unpublished Doctoral Dissertation, Portland State University, Portland.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). Boston: Kluwer Academic Publishers.

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

Creswell, J. W. (1994). *Research Design: Qualitative and Quantitative Approaches*. Thousand Oaks: SAGE.

Dabos, M. G. (2009). *Two-Year College Mathematics Instructors' Conceptions of Variation* (Unpublished master's project). University of California Santa Barbara: Santa Barbara, CA


.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55-82. [Online: www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_delMas_Liu.pdf]

delMas, R. C., & Liu, Y. (2007). Students' conceptual understanding of the standard deviation. In M. C. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 87-116). New York: Lawrence Erlbaum Associates.

Delucchi, M. (2007). Assessing the impact of group projects on examination performance in social statistics. *Teaching in Higher Education, 12*(4), 447-460.

Erickson, T. (July, 2006). *Using simulation to learn about inference.* Paper presented at the Seventh International Conference on Teaching Statistics, Salvador, Brazil.

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: towards an assessment framework [Electronic Version]. *Journal of Statistics Education*, 2. Retrieved 11/24/2005 from www.amstat.org/publications/jse/v2n2/gal.html.

Gardner, P. L., & Hudson, I. (1999). University students' ability to apply statistical procedures [Electronic Version]. *Journal of Statistics Education*, 7. Retrieved 8/29/2006 from www.amstat.org/publications/jse/v7n1/gardner.html.

Garfield, J. (2002). The challenge of developing statistical reasoning [Electronic Version]. *Journal of Statistics Education*, 10. Retrieved March 27, 2006 from http://www.amstat.org/publications/jse/V10N3/garfield.html.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44 - 61.

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

Garfield, J., delMas, R., & Chance, B. (2002). *The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project*. NSF CCLI grant ASA-0206571. [Retrieved from https://app.gen.umn.edu/artist/]

Garfield, J., delMas, R. C., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In

M. C. Lovett & P. Shah (Eds.), *Thinking With Data* (pp. 117-147). New York: Lawrence Erlbaum Associates.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First course in statistical science: The status of educational reforms efforts [Electronic Version]. Journal of Statistics Education, 10. Retrieved 8/15/07 from http://www.amstat.org/publications/jse/v10n2/garfield.html#garfield2000.

Johnson, R. B. (2006). Data Analysis in Qualitative Research: University of South Alabama.

Kirkman, E., Lutzer, D., Maxwell, J. W., & Rodi, S. B. (2007). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2005 CBMS Survey*. Washington DC: Conference Board of Mathematical Sciences.

Konold, C., & Higgins, T. (2002). Highlights of related research. In S. J. Russell, D. Schifter & V. Bastable (Eds.), *Developing mathematical ideas: Working with data* (pp. 165-201). Parsippany: Dale Seymore Publications.

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

Konold, C., & Pollatsek, A. (2004). Conceptualizing an average as a stable feature of noisy process. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 169-200). Boston, MA: Kluwer Academic Publishers.

Lavigne, N. C., & Glaser, R. (2001). *Assessing student representations of inferential statistics problems*. Los Angeles: University of California Los Angeles.

Leavy, A. (2006). Using data comparisons to support a focus on distribution: Examining preservice teachers' understandings of distribution when engaged in statistical inquiry [Electronic Version]. *Statistics Education Research Journal*, 5, 89-114. Retrieved July 15, 2007, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Leavy.pdf.

Lutzer, D., Maxwell, J. W., & Rodi, S. B. (2002). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2000 CBMS Survey*. Providence, RI: American Mathematical Society.

Makar, K., & Rubin, A. (2007). *Beyond the bar graph: Teaching informal statistical inference in primary school*. Paper presented at the Conference Name|. Retrieved Access Date|. from URL|.

Masnick, A. M., Klahr, D., & Morris, B. J. (2007). Separating signal from noise: Children's understanding of error and variability. In M. C. Lovett & P. Shah (Eds.), *Thinking with Data* (pp. 3-26). New York: Lawrence Erlbaum Associates.

Meletiou-Mavrotheris, M., & Lee, C. (February, 2005). *Exploring introductory statistics students' understanding of variation in histograms.* Paper presented at the Fourth Congress of ERME, the European Society for Research in Mathematics Education, Sant Feliu de Guíxols, Spain.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants* (pp. 95–138). Washington, D.C.: National Academy Press.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123-137.

Moore, D. S., & McCabe, G. P. (1993). *Introduction to the Practice of Statistics* (2nd ed.). New York: Freeman.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Noll, J. (2007). *Graduate teaching assistants' statistical knowledge for teaching.* Unpublished Doctoral Dissertation, Portland State University, Portland

Patton, M.Q. (2002). *Qualitative Research and Evaluation Methods.* Thousand Oaks, CA: Sage.

Peck, R. (2005). *There's more to statistics than computation - Teaching students how to coummunicate statistical results.* Paper presented at the Statistics Education and the Communication of Statistics, Sydney, Australia.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27-45.

Pfannkuch, M., & Reading, C. (2006). Reasoning about distribution: A complex process [Electronic Version]. *Statistical Education Research Journal*, 5, 4-9. Retrieved July 15, 2007 from http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2).pdf.

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

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistical Education Research Journal, 7*(2), 107-129.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88*(1), 144-161.

Quilici, J.L. & Mayer, R.E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology, 16,* 325-342.

Reading, C., & Reid, J. (2004). Consideration of variation: A model for curriculum development, presented at the *International Association for Statistical Education Roundtable on Curricular Development in Statistics Education*, Lund, Sweden, 28 June to 3 July.

Reading, C., & Reid, J. (2005). Consideration of variation: A model for curriculum development. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable* (pp. 36-53). Voorburg, The Netherlands: International Statistical Institute.

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

Reading, C., & Reid, J. (July, 2006). *Listen to the students: Understanding and supporting students' reasoning about variation.* Paper presented at the Seventh International Conference on Teaching Statistics, Salvador, Brazil.

Reading, C., & Reid, J. (2007). Reasoning about variation: Student voice. *International Electronic Journal of Mathematics Education, 2*(3), 110-127.

Reading, C., & Shaughnessy, J.M. (July, 2000). *Students perceptions of variation in a sampling situation.* Paper presented at the 24th Conference of the International Group for the Psychology of Mathematics Education, Hiroshima, Japan.

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.

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistical Education Research Journal, 7*(2), 5-19.

Rubin, A., Bruce, B., & Tenney, Y. (August, 1990). *Learning about sampling: Trouble at the core of statistics.* Paper presented at the International Conference on Teaching of Statistics (ICOTS-3), Dunedin, New Zealand.

Saldanha, L. A., & Thompson, P. W. (2007). Exploring connections between sampling distributions and statistical inference: An analysis of students' engagement and thinking in the context of instruction involving repeated sampling. *International Electronic Journal of Mathematics Education, 2*(3), 270-297.

Shaughnessy, J.M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph and K. Carr (Eds.), *People in mathematics education* (Vol. 1, pp.6-22). Waikato, New Zealand: Mathematics Education Research Group of Australasia.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester Jr. (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (Vol. 2, pp. 957-1009). Reston: National Council of Teachers of Mathematics.

Shaughnessy, J. M., Ciancetta, M., & Canada, D. (July, 2004). *Types of student reasoning on sampling tasks.* Paper presented at the 28th Conference of the International Group for the Psychology of Mathematics Education, Bergen, Norway.

Shaughnessy, J. M., Ciancetta, M., Best, K., & Canada, D. (April, 2004). *Students' attention to variability when comparing distributions.* Paper presented at the 82nd Annual Meeting of the National Council of Teachers of Mathematics, Philadelphia, PA.

Shulman, L. S. (1986). Those who understand knowledge growth in teaching. *Educational Researcher, 15*(2), 4 -14.

Sorto, M. A. (2004). *Prospective middle school teachers' knowledge about data analysis and its application to teaching.* Michigan State University, East Lansing.

Sotos, A., Vanhoof, S., den Noortgate, W., and Onghena, P. (2009) How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education* 17(2).
Retrieved from: http://www.amstat.org/publications/jse/v17n2/castrosotos.html

Thompson, P. W., Liu, Y., & Saldanha, L. A. (2007). Intricacies of statistical inference and teachers' understanding of them. In M. C. Lovett & P. Shah (Eds.), *Thinking With Data* (pp. 207-232). New York: Lawrence Erlbaum Associates.

Tremblay, P. F., Gardner, R. C., & Heipel, G. (2000). A model of the relationships among measures of affect, aptitude, and performance in introductory statistics. *Canadian Journal of Behavioral Science, 32*(1), 40-48.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think-aloud method: A practical guide to modeling cognitive processes*. San Diego, CA: Academic Press Ltd.

Watson, J. M. (2007). The role of cognitive conflict in developing students' understanding of average. *Educational Studies in Mathematics, 65*(1), 21-47.

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

Watson, J. M., & Kelly, B. A. (2002.). Emerging concepts in chance and data. *Australian Journal of Early Childhood, 27*(4), 24-28.

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.

Wild, C., & Pfannkuch, M. (1998). What is statistical thinking? In L. Pereira-Mendoza, L. Kea, T. Kee & W. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (Vol. 1, pp. 333-339). Voorburg: ISI.

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

Yilmaz, M. (1996). The challenge of teaching statistics to non-specialists [Electronic Version]. *Journal of statistics education*, 4. Retrieved 6/05/06 from http://www.amstat.org/publications/jse/v4n1/yilmaz.html.

Zieffler, A., Garfield, J., DelMas, R. C., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistical Education Research Journal, 7*(2), 40-58.

## Appendix A

### Instrument A

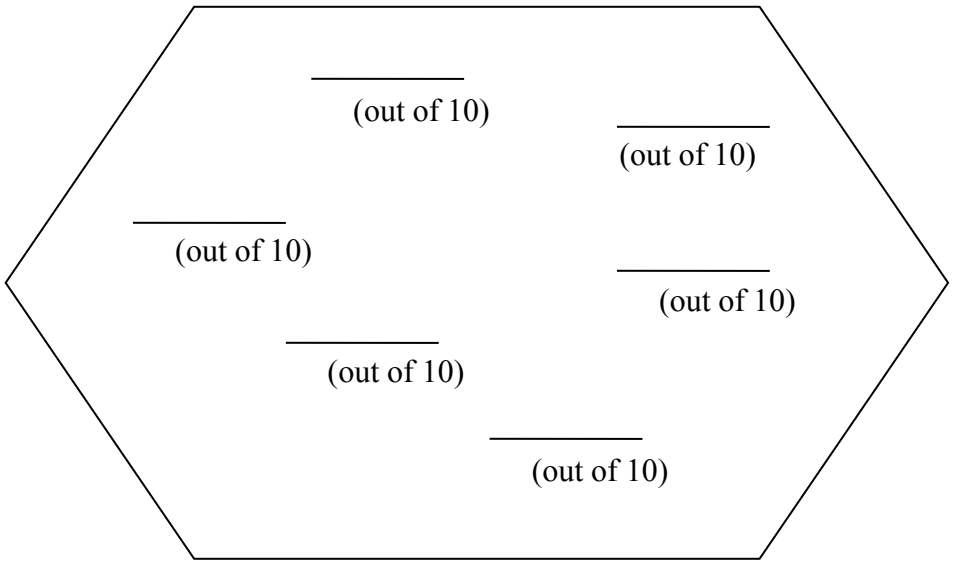
This set of questions help to give a picture of how you think about some problems in probability and statistics. Rather than think of the right or wrong answer, just write down your best thinking for the situation.

1] Suppose there is a container with 100 pieces of candy in it. 60 are Red, and 40 are Yellow. The candies are all mixed up in the container.You reach in and pull out a handful of 10 candies at random.

(a) How many red candies do you think you might get?
Why do you think this?

(b) Suppose you do this several times (each time returning the previous handful of 10 candies and remixing the container). Do you think this many reds would come out every time? Why do you think this?

(c) Suppose six students do this experiment (each time returning the previous handful of 10 candies and remixing the container). Write down the number of reds that you think each student might get:




Why did you choose those numbers?

**2]** Suppose 6 people did this experiment – pulled ten candies form the container, wrote down the number of reds, then returned the ten and remixed all the candies.

a) What do you think the number of reds will most likely to go from?

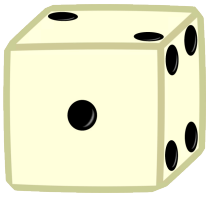
From a low of _____ to a high of _____.

b) Now suppose 30 people did this experiment. What do you think the number of reds will most likely go from?

From a low of _____ to a high of_____

c) Why do you think this?

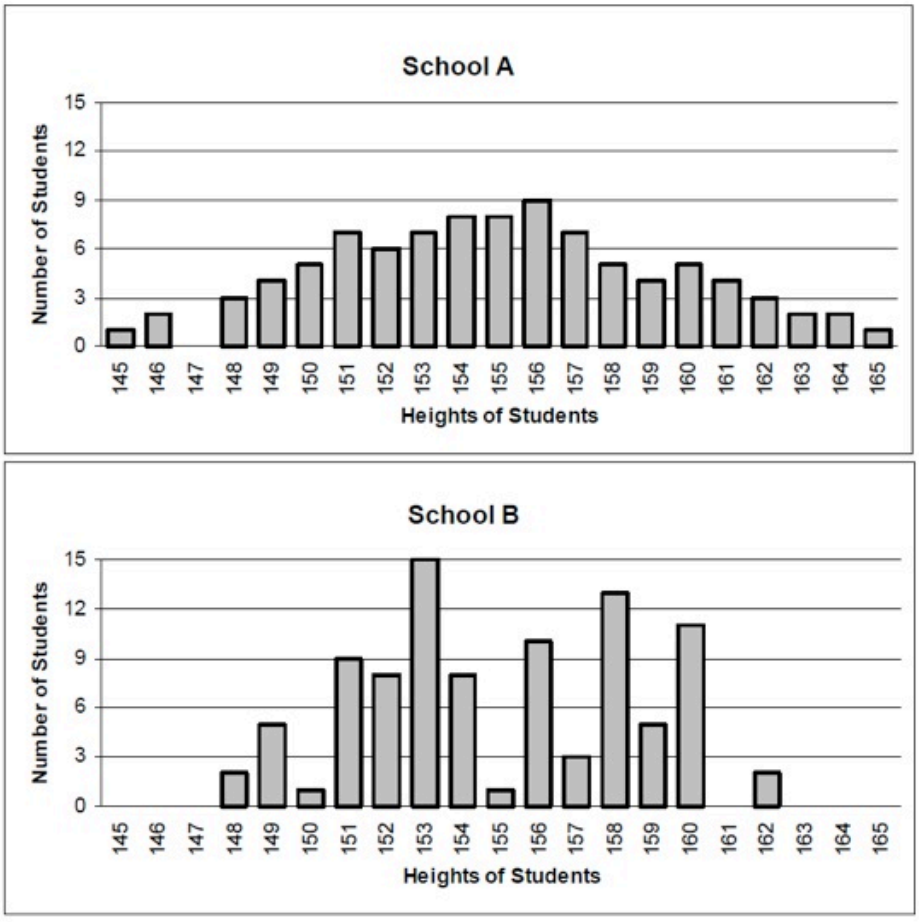
**3] Consider a regular, fair, six-sided die. Imagine that you threw the die 60 times. Fill in the table below to show how many times you think each number might come up.**



| Number that shows on the tossed die | How many times it might come up? |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| Total = | 60 |

**Why do you think that this numbers are reasonable?**

**4]** The following graphs describe some data collected about Grade 7 students' heights (measured in centimeters) in two different schools:



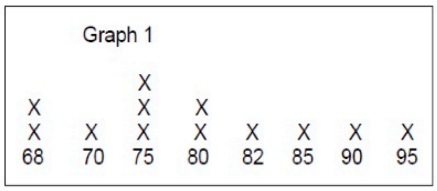

**a) Which graph shows more variability in students' heights?**

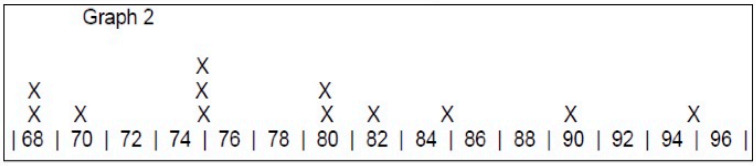
**b) Explain why you think this?**

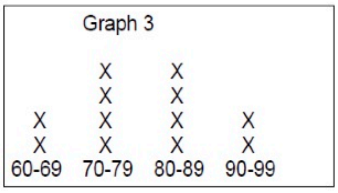
**5]** A new car was being tested to see how well the brakes worked. The test engineer measured how many inches the car took to slow from 40 mph to 0 mph; the fewer inches taken, the better the braking power. Twelve trials were run, under the same road conditions and with the same test driver. Here were the results ( to the nearest inch):

| Stopping Distance (in) | | | |
|---|---|---|---|
| 68 | 68 | 70 | 75 |
| 75 | 75 | 80 | 80 |
| 82 | 85 | 90 | 95 |

The engineer was then trying to decide how to graph the results. She came up with the following three graphs for representing the data:



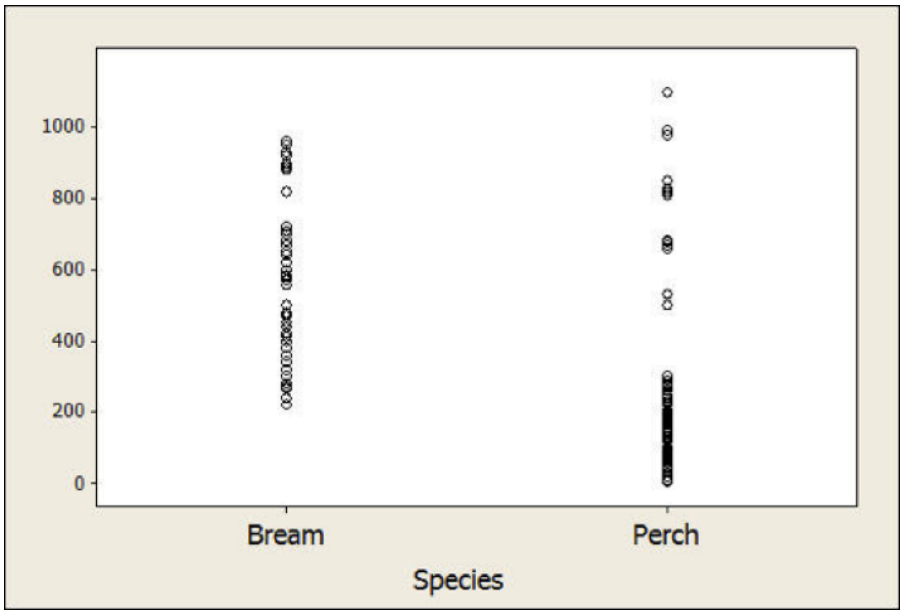

a) Do these graphs differ in the way they show the braking power? If so, how?

b) Do you think one graph shows more variability in the results than others? Explain.

c) If the engineer wanted to suggest that the car was fairly consistent in its braking power, which graph would you suggest she use, and why?

d) Does one graph help you more than the other in making your conclusion?

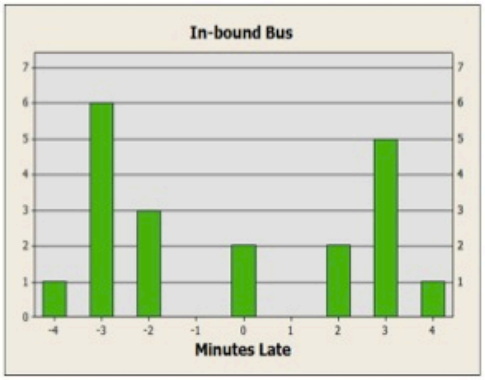
**6]** Look at the following plot. It shows the weights in grams of two species of fish (bream and perch).

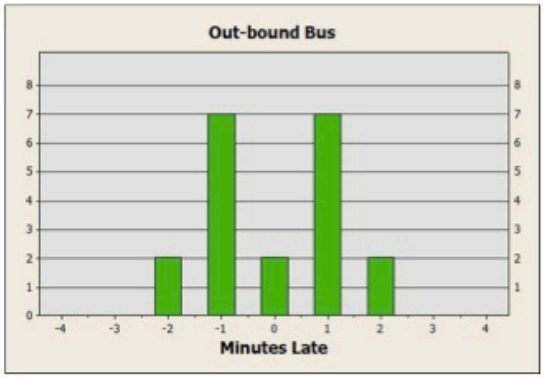



a) Do you think that there is a difference in weight for the two species?

Explain your response.

***7]*** *Citizens in an outer suburb were concerned about the reliability of their bus service to the center of the city. They monitored the in-bound and out-bound service of the buses at Bus Stop 33, and recorded the number of minutes late. Zero minutes late indicates the bus was on time while a negative number of minutes late indicates the bus was early. The data are displayed in the two graphs. Describe and compare the performances of the two bus routes.*



**8]** At a nearby college, half the students are women and half are men. A worker for a student organization wants to interview students on their views about recent changes in the federal government's funding of financial aid. The worker wants to get a good representation of the students, and goes to as many different areas on campus as possible. Three or four students are interviewed at each place the worker visits. Out of the last 20 students interviewed, 13 were women and 7 were men. Now, you do not know what time of day it is, to which part of campus the worker has already gone, or where the worker is going next. Out of the next 20 students the worker interviews, do you think more will be women or men?

**a.** The worker seems to interview more women than men. There could be several reasons for this. Perhaps women are more willing to talk about their opinions. Or, maybe the worker goes to areas of campus where there more women than men. Either way, the worker is likely to interview more women than men out of the next 20 students.

**b.** Since half of the students on this campus are men and half are women, you would expect a 50/50 split between the number of men and women the worker interviewed. Since there tended to be more women than men so far, I expect the opposite trend to start happening. Out of the next 20 students the worker interviews, there will probably be more men than women so that things start to balance out.

**c.** Half the students on this campus are men and half are women. That means that the worker has a 50/50 chance of interviewing a man or a woman. It should not matter how many men or women the worker has interviewed so far. Out of the next 20 students interviewed, about half should be men and half women.

**9]** Nine students in a science class weighed a small object on the same scales separately. The weights (in grams) recorded by each student are shown below.

6.2 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.2

The students want to determine as accurately as they can the actual weight of this object. They may use the following methods:

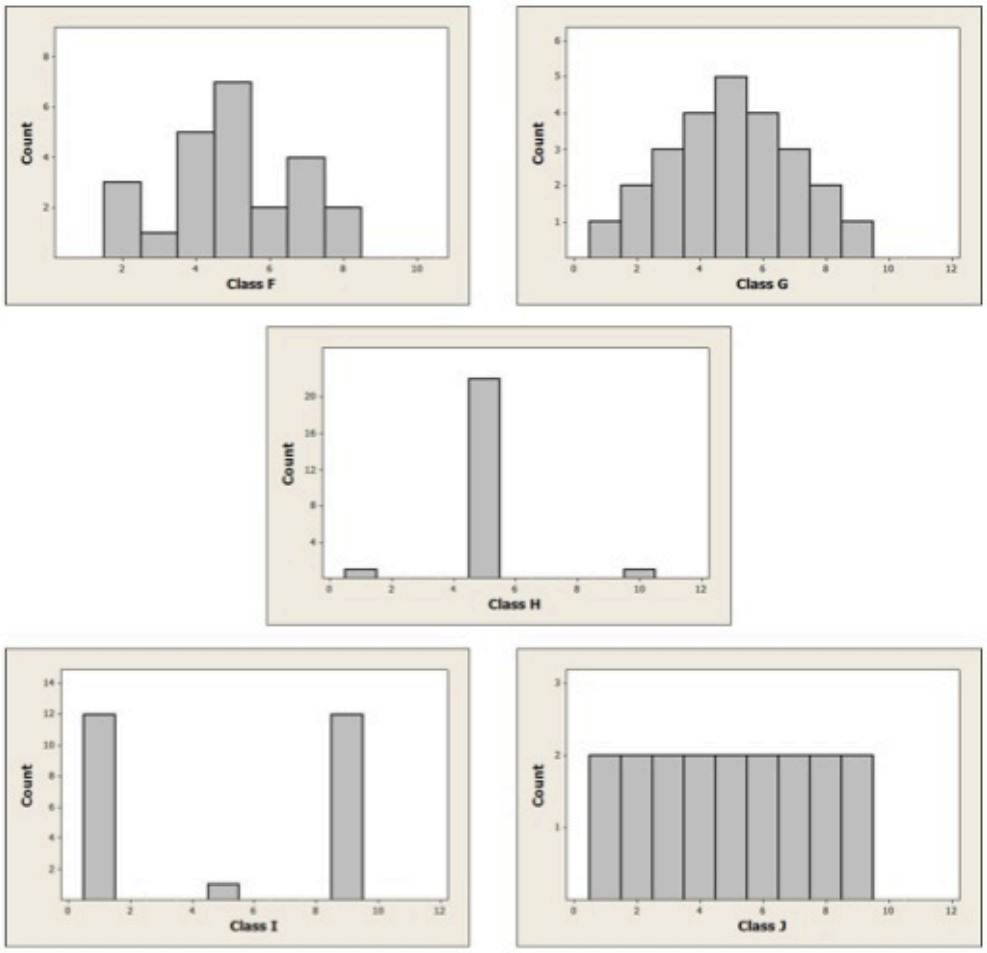
I. Use the most common number, which is 6.2
II. Use the 6.15 since it is the most accurate weighing.
III. Use the result of adding up the 9 numbers and dividing by 9.

As teacher, what method would prefer your students use?

a) Method I
b) Method II
c) Method III
d) Other_____

Explain your choice.

***10]*** *Suppose that students in five hypothetical statistics classes (Class F, Class G, Glass H, Class I, Class J) were asked to rate the value of statistics on a 1-9 scale. The ratings of each of the classes are shown in the following histograms.*




a) Judging from the histograms, take a guess as to which has more variability between classes F and G. Explain

b) Judging from the tables and histograms, which would you say has the most variability among class H, I, and J? Which would you say has the least variability? Explain

## Appendix B

## INSTRUMENT B

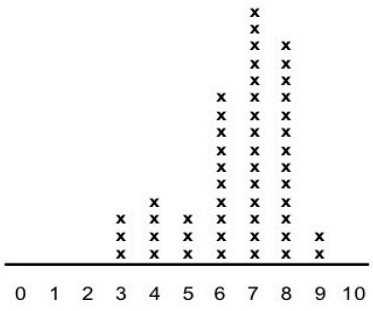
## THINK ALOUD

**Please solve the following problems and while you do so, try to say everything that goes through your mind.**

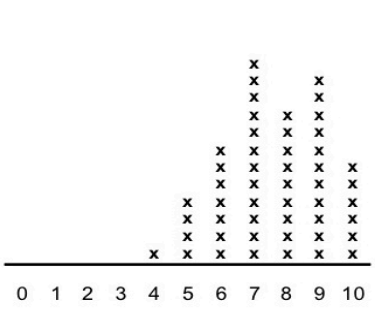
**1]** Real/ Fake

A class conducted an experiment, pulling 50 samples of 10 candies from a jar with 750 red and 250 yellow, 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. Which graphs do you think are real? Which graphs do you think are made-up. Explain your reasons for your choices.

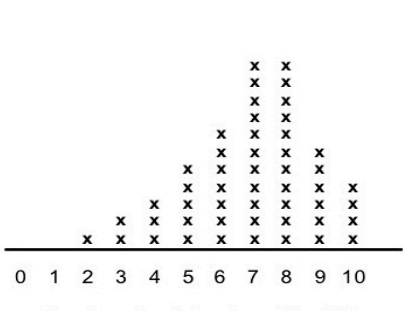
**(Graph 1)**



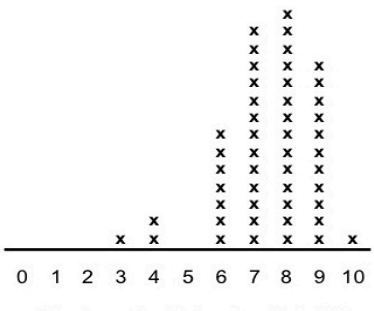

**(Graph 2)**




**(Graph 3)**




**(Graph 4)**



**2]** A bowl has 100 wrapped hard candies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Jenny pulls out a handful of 10 candies while blindfolded, counts the number of reds, and tells her teacher. The teacher writes the number of red candies on a list. Then, Jenny puts the candies back into the bowl, and mixes them all up again. Five of Jenny's classmates, Jack, Julie, Jason, Jane and Jerry do the same thing. They each pick ten candies count the reds, and the teacher writes down the number of reds. Then they put the candies back and mix them up again each time.

a) What do you think the teacher's list for the number of reds is likely to be? Explain why you chose those numbers.

b) If you were asked to choose a response to this question from the following list, circle the one that you would choose. Explain why you chose that one.

A) 5, 9, 7, 6, 8, 7  B) 3, 7, 5, 8, 5, 4  C) 5, 5, 5, 5, 5, 5  D) 2, 4, 3, 4, 3, 4  E) 3, 0, 9, 2, 8, 5

**INSTRUMENT B**
**THINK ALOUD**

**Please solve the following problems and while you do so, try to say everything that goes through your mind.**

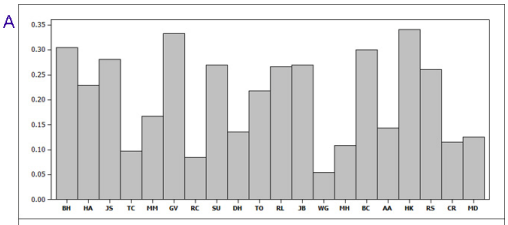
**3]** A baseball fan likes to keep track of statistics for the local high school baseball team. One of the statistics she recorded is the proportion of hits obtained by each player based on the number of times at bat as shown in the table below. Which of the following graphs gives the best display of the distribution of proportion of hits in that it allows the baseball fan to describe the shape, center and spread of the variable, proportion of hits?

| Player | Proportion of hits |
|---|---|
| BH | 0.305 |
| HA | 0.229 |
| JS | 0.281 |
| TC | 0.097 |
| MM | 0.167 |
| GV | 0.333 |
| RC | 0.085 |

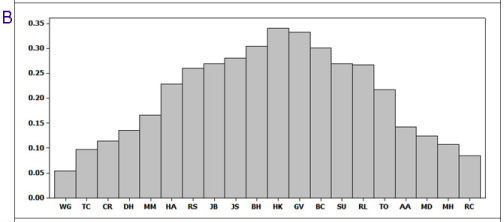
| Player | Proportion of hits |
|---|---|
| SU | 0.270 |
| DH | 0.136 |
| TO | 0.218 |
| RL | 0.267 |
| JB | 0.270 |
| WG | 0.054 |
| MH | 0.108 |

| Player | Proportion of hits |
|---|---|
| BC | 0.301 |
| AA | 0.143 |
| HK | 0.341 |
| RS | 0.261 |
| CR | 0.115 |
| MD | 0.125 |

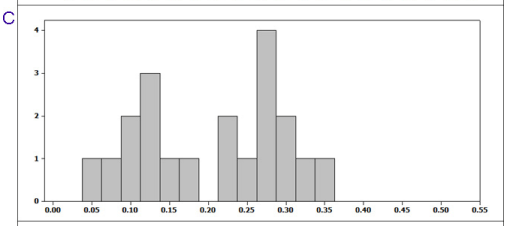
A



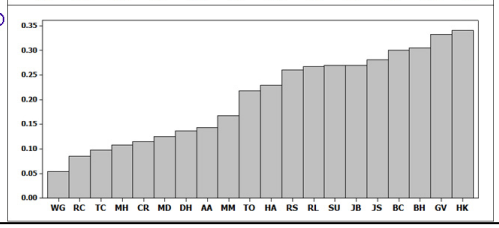

B



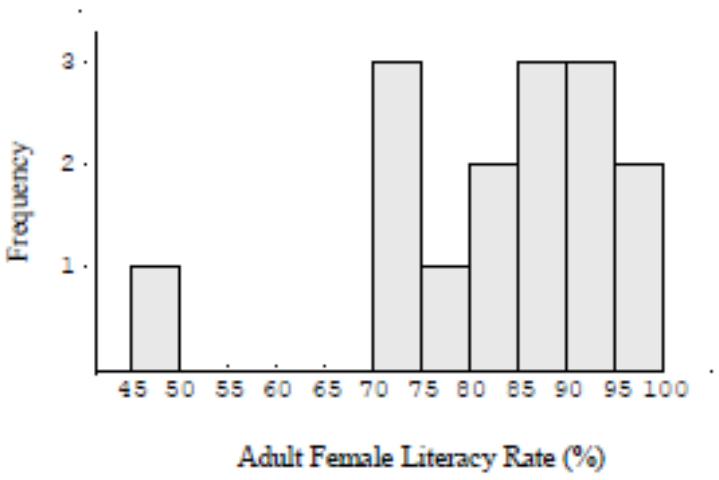

C




D



4] The following graph gives information about adult female literacy rates in Central and South American countries,




Adult Female Literacy Rates in Central and South America

a. Suppose you ask your students to tell you how many countries are represented in the graph? One student says, "there are 7 countries represented".

Is this student right or wrong? _____

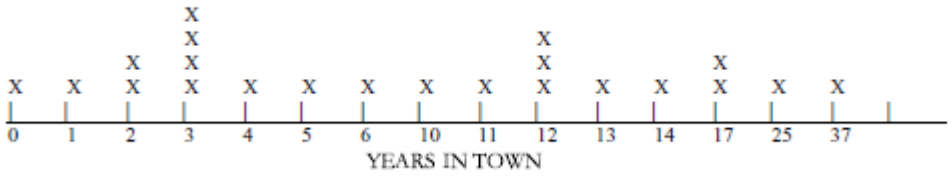
In your opinion, what is the student's thinking to arrive to that conclusion?

b. Suppose now you ask your students to explain what the third bar from the right indicates. One says, "It indicates 85% to 90% literacy rate".

Comment on the response

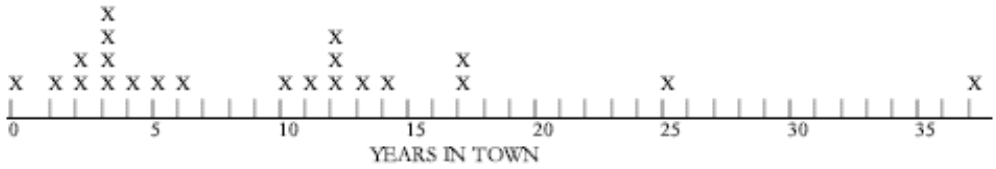
**5]** A class of students recorded the number of years their families had lived in their town. Here are two graphs that students drew to tell the story.

**Graph 1**




a) What can you tell by looking at Graph 1?

**Graph 2**




b) What can you tell by looking at Graph 2?

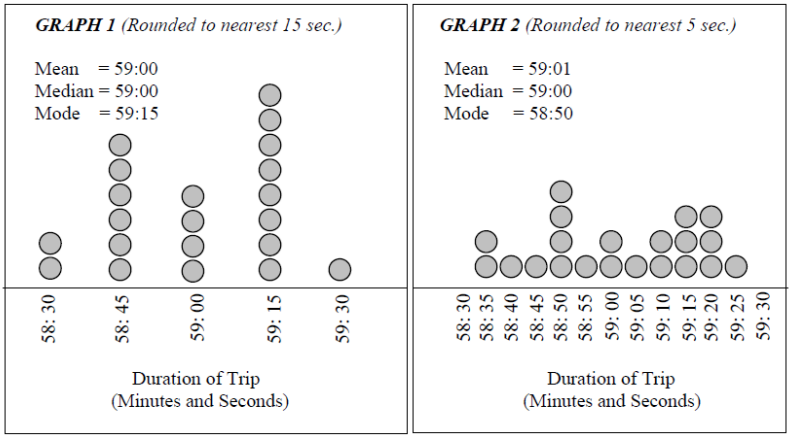
c) Which of these graphs tells the story better? Why

**6]** A class of twenty-one students wanted to find out some information about MAX train rides. Their first goal was to find out the duration of a ride from Washington Park to Gresham. They all got on the same train, but they sat separately and kept track of the time on their own. Later in class, they were surprised to find that they did not have the same results:

| Duration of Ride ( Min:Sec , to the nearest second) | | | | |
|---|---|---|---|---|
| 58: 36 | 58 :36 | 58: 40 | 58: 44 | 58: 51 |
| 58: 50 | 58: 49 | 58: 50 | 58 :56 | 59: 01 |
| 59: 02 | 59: 06 | 59: 11 | 59: 09 | 59: 16 |
| 59 :14 | 59: 15 | 59: 19 | 59: 21 | 59: 20 |
| 59: 24 | | | | |

a) What are some possible reasons for why the class did not get the same result?

The Class was deciding how to display this data. In Graph I, they rounded to the nearest 15 seconds. In graph 2, they rounded to the nearest 5 seconds.




a) How do these graphs differ in their stories they tell about the duration of the trip?

b) Some members of the class argue that the trip was really under 59 minutes, while some argue that it was over 59 minutes. Others claim it was exactly 59 minutes. What do you think about the true duration of the trip, and why do you think this?

# Appendix C

## Informed Consent

You are invited to participate in a doctoral research project entitled "Introductory Statistics Professors' Conceptions of Variation", being conducted by Monica Dabos from the Department of Education at the University of California Santa Barbara. Through this research I hope to describe the knowledge (intuitive or learned) held by introductory statistics professors at two year colleges about the important statistical concept of variation. You were selected as a possible participant by virtue of your experience as an instructor or tutor in the mathematics department at the two year college level.

If you choose to participate in this project, I will collect two kinds of data. 1) I will ask you to answer a written instrument. The instrument consists of 10 questions that can be answered in approximately half an hour. 2) Then I will give you a set of 6 more questions; during this section, I will ask you to explain what you are doing while you do it. The entire data collection will take about 1 hour and will take place in a reserved room in the library at SBCC or at your chosen location. While you answer both parts there will be a voice recording and a videotape recording in the room.

You, as a two year college instructor, may gain a direct benefit from a deeper exploration of your own ideas about this key statistical concept; this exploration allows you to extend your own learning about variation in a non-evaluative environment. Moreover, the practice of articulating your thinking may be especially helpful as you discuss similar concepts with your own students. In the future this research may help other instructors to better teach the concepts of variation to statistics students.

Potential risks include the possibility that an unauthorized person may view the data, or that your actual name may inadvertently become associated with the data. To minimize this risk, all written responses, notes, audio and video tapes, and transcriptions will be kept confidential, and will be kept locked up in the researcher's office in the Department of Education at UCSB. The only people who will see the data are myself and my faculty advisors. In writing any results for the study, pseudonyms will be used so that your identity cannot be matched with the responses you have provided. There is also a risk that having a researcher with a statistics background conducting the study may make you feel uncomfortable. To lessen this risk, I want you to know that this research is descriptive and not evaluative in nature.

Your participation in this study is voluntary and you are completely free to withdraw from the study at any time. Your decision to participate or not will not affect your relationship with the researcher or with any academic program at UCSB or at your institution in any way. If you have any questions about the study itself, please contact Monica Dabos, at The Gevirtz School, UC Santa Barbara (805) 452-2107. Your signature indicates that you have read and understand the above information and agree to take part in this study. Please remember that you may withdraw your consent at any time without penalty.

_____ _____

Signature of Participant Date

PLEASE KEEP A COPY FOR YOUR RECORDS

Questions or problems about your rights in this research project can be directed to Kathy Graham; Human Subjects Committee; Office of Research; 3227 Cheadle Hall, University of California Santa Barbara. Santa Barbara, CA 93106. Telephone: (805) 893 3807