CHARACTERIZING MIDDLE AND SECONDARY PRESERVICE TEACHERS' CHANGE IN INFERENTIAL REASONING

A Dissertation

presented to

the Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

MARYANN E. HUEY

Dr. James E. Tarr, Dissertation Supervisor

MAY 2011

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled

CHARACTERIZING MIDDLE AND SECONDARY PRESERVICE TEACHERS'

CHANGE IN INFERENTIAL REASONING

Presented by Maryann E. Huey

A candidate for the degree of Doctor of Philosophy

And hereby certify that in their opinion is worthy of acceptance.

_____
James E. Tarr

_____
David A. Bergin

_____
Óscar Chávez

_____
Kathryn B. Chval

_____
Douglas A. Grouws

For Elise and Bryce

ACKNOWLEDGEMENTS

If one were knowingly to solicit the necessary support required to complete a dissertation in advance, few would have the courage to make such a large request of faculty, family and friends. Therefore, as I reflect on the countless supports provided to me, I view this document as not just a reflection of my efforts, but also of those who have encouraged me, shaped my thinking and provided ongoing inspiration.

I would like to thank Dr. James Tarr for his commitment to producing a rigorous research study, dedication during the writing and revising stages, and expertise in statistics education. Dr. Tarr removed many potential obstacles to completing this project and set high expectations for my work. This combination allowed me to focus on key elements of the study in a supportive environment and to truly enjoy all aspects of completing this dissertation. I am also grateful for the support and insight provided by the members of my committee: Dr. David Bergin, Dr. Óscar Chávez, Dr. Kathryn Chval, and Dr. Douglas Grouws. The committee's engagement in framing the research study strengthened the final product significantly.

I would also like to thank Dr. Barbara Reys for the financial support provided in order to present portions of my dissertation work at national conferences. These experiences allowed me to interact face-to-face with leaders in the statistics education community and receive additional perspectives on my work. As a result, I gained a broader view of statistics education research and a deeper understanding of how my dissertation study could address needs within the field. In addition, I sincerely appreciate the instructor's willingness to open the statistics course to my research efforts and

encourage the preservice teachers to participate. Finally, I would like to acknowledge the technical expertise provided by Chris Bowling in order to improve the quality of scanned images contained within this document.

Family, friends and colleagues form the foundation of our lives and make all dreams possible. Both my mother, Beverly Baartmans, and father, Gary Gimmestad, have always assured me that I could achieve any goal I desired through their frank acceptance of my most ambitious plans and their own lives as models. My sister, Katherine Gimmestad, has been my comrade and confidant, as we have both begun and completed dissertations within the state of Missouri during similar timeframes. Friends and colleagues, Anne Estapa, Mary Turner, Leslie Adrian, Dr. Cynthia Taylor, Trish Green, Dr. Sarah Hicks, and members of St. Andrew's Lutheran Church, have made daily work enjoyable and shared in accomplishments along the way.

Of all the people mentioned, I credit my husband, Brian, the most. He agreed to move from Kansas City, alter his own career, commute long distances, and share in all family responsibilities to ensure my successful completion of this dissertation. Brian's enthusiasm never waned throughout graduate school, and he has maintained an inspirational focus on the future. My children, Elise and Bryce, remind me continuously that we owe our children the very best we can offer, as they give their very best each and every day.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS……...……...ii

LIST OF TABLES…….x

LIST OF FIGURES……..xi

ABSTRACT…….xiv

Chapter

1. STATEMENT OF THE PROBLEM AND BACKGROUND.....1

The Purpose of the Study.....1

Teachers' Statistical Reasoning.....4

Inferential Reasoning.....6

Developing Inferential Reasoning Abilities.....7

Reasoning Summary.....8

Conceptual Framework.....8

Stage 1: Informal Inferential Reasoning Components.....9

Stage 2: Formal Inferential Reasoning Components 10

Theoretical Framework.....11

The Structure of Observed Learning Outcomes (SOLO) Taxonomy.....11

Characterization of Informal Inferential Reasoning Responses.....14

Characterization of Formal Inferential Reasoning Responses.....15

Response Classification.....15

Reasoning Progression.....15

Significance of the Study....16

Summary....18

2. LITERATURE REVIEW....20

The Rise of Inferential Reasoning in the Mathematics Curriculum....20

The Evolution of Statistics Standards....21

Endorsing Reasoning and Sense-Making....24

Inferential Reasoning....25

The Process of Inferential Reasoning....26

Data Distributions....27

Sample and Sampling Variation....30

Variance....32

Inferential Reasoning Summary....33

Teachers' Understanding of Inferential Reasoning....34

Characteristics of Tasks Used to Assess Inferential Reasoning....38

Ill-Structured....38

Open-Ended....40

Visual Representations....41

Context....42

Summary....44

3. RESEARCH DESIGN AND METHODOLOGY....47

Research Methodology....47

Research Context....47

Participants....49
Subject selection....49
Background coursework....50
Accessing participants....51
Data Sources....51
Designing the assessments....53
Pre-assessment....58
Midcourse clinical interviews....59
Post-assessment....60
Artifacts, audio recordings and field notes....61
Data Collection....61
Data Analysis....62
Characterization of Informal Inferential Reasoning Responses....63
Application of the informal inference framework....64
Characterization of Formal Inferential Reasoning Responses....66
Application of the formal inference framework....67
Dominant Level of Reasoning....70
Informal and Formal Inferential Reasoning....72
Opportunity to Learn....72
Reliability....75
Limitations of the Study....76
Summary....76

4. ANALYSIS OF THE DATA AND RESULTS 78

Characterizing the Change in the Preservice Teachers' Inferential Reasoning.....78

Variation in Individual Reasoning Responses at the Class Level.....79

Instability of Inferential Reasoning Across Tasks.....86

Pre-assessment.....86

Post-assessment.....88

Stability in Inferential Reasoning Across Time.....92

Stable Unistructural thinking.....92

Stable Multistructural thinking.....101

Growth in Inferential Reasoning Over Time.....108

Unistructural to Multistructural.....108

Prestructural to Unistructural 111

Multistructural to Relational.....118

Decline in Inferential Reasoning Over Time.....122

Unistructural to Prestructural.....122

Relation to Prior Statistics Coursework.....128

Characterization of Formal Inferential Reasoning.....129

Variation in Formal Inferential Reasoning Responses at the Class Level.....130

Common Problematic Features of Formal Inferential Reasoning.....134

Association of Informal and Formal Inferential Reasoning.....136

Opportunity to Learn.....138

Mathematical Strands of Proficiency.....141

Textbook.....141

Computer Simulations.....142

Quantitative Literacy Series.....143

Instructor-generated.....143

Exams.....144

Chronological view.....144

Content Analysis.....146

Core concepts.....150

Aggregate concepts.....152

Results Summary.....154

5. DISCUSSION, SUMMARY, AND RECOMMENDATIONS 156

Summary of the Study and Findings.....157

Methodology.....157

Results of the Study.....159

Preservice middle and secondary mathematics teachers' inferential reasoning.....159

Relationship between informal and formal inferential reasoning.....163

The opportunity to learn inferential Reasoning.....163

Discussion of Findings.....165

Limitations of the Study.....171

Nature of the Assessment Tasks.....171

Formal Inference Characterization.....172

Implications for the Statistical Preparation of Teachers.....173

Greater Emphasis on Adaptive Reasoning.....173

Explicit Focus on Proportional Reasoning.....174

The Role of Probability: A Tool for Statistics.....174

Explicit Attention to Informal Inferential Reasoning.....175

Recommendations for Future Research.....176

The Development of Inferential Reasoning.....177

Formal Inference Characterization.....178

The Relationship between Informal and Formal Inferential Reasoning.....178

Context and Data Representations of Tasks.....179

Reflections.....179

REFERENCES.....182

APPENDIX A: PRE-ASSESSMENT.....188

APPENDIX B: MIDCOURSE ASSESSMENT.....199

APPENDIX C: POST-ASSESSMENT.....207

VITA.....219

LIST OF TABLES

Table		Page
3.1	Participants' Academic Standing and Gender Profile	50
3.2	Prior Tertiary Statistics Coursework Completed	51
3.3	Data Sources: Timeline, Frequency and Purpose	52
3.4	Task Summary for the Assessments	58
3.5	Informal Inference Characterization Framework	63
3.6	Formal Inference Characterization Framework	67
4.1	Dominant Levels of Inferential Reasoning on the Pre-Assessment	88
4.2	Dominant Levels of Inferential Reasoning on the Post-Assessment	90
4.3	Shifts in Levels of Inferential Reasoning, Pre- to Post-Assessment	92
4.4	Proportion of the Course Dedicated to Each Section of Content	139
4.5	Proportion of the Tasks and Codes Generated from Each Source	141

LIST OF FIGURES

Figure Page

1.1 Conceptual framework for inferential reasoning development……...9

1.2 The SOLO model……...13

2.1 Timeline of shifting emphases on statistics and probability in the K-12 school curriculum……...21

3.1 Sample item, Ambulance Service task……...55

3.2 Hypothetical informal responses to the diet and cholesterol task……..65

3.3 Hypothetical formal responses to the hiring discrimination task……...69

3.4 Strands of mathematical proficiency from *Adding It Up*……..……74

4.1 Example of a Unistructural response……82

4.2 Example of a Multistructural response……..83

4.3 Example of a Relational response……..85

4.4 Middle and secondary mathematics preservice teachers' combined assessment results……..91

4.5 Unistructural response to the Migraine Treatment task on the pre-assessment……...94

4.6 Unistructural response to the Ambulance Service task on the midcourse assessment……95

4.6 Unistructural to Multistructural response for the Speed Trap task on the midcourse assessment……96

4.7 Unistructural response to the Migraine Treatment task on the post-assessment…….……99

4.8 Unistructural response to the Diet and Cholesterol task on the post-assessment……100

4.9 Multistructural response to the Diet and Cholesterol task on the pre-

assessment……102

4.10 Multistructural response to the Ambulance Service task on the midcourse assessment……..……103

4.11 Relational response to the Speed Trap task on the midcourse assessment……105

4.12 Multistructural response to the Diet and Cholesterol task on the post-assessment……107

4.13 A Unistructural response to the Diet and Cholesterol task on the pre-assessment……109

4.14 A Multistructural response to the Diet and Cholesterol task on the post-assessment……110

4.15 A Prestructural response to the Class Scores task on the pre-assessment…....112

4.16 A Prestructural response to the Training Programs task on the pre-assessment……...113

4.17 A Unistructural response to the Migraine Treatment task on the post-assessment……115

4.18 A Unistructural response to the Speed Trap task on the midcourse assessment……116

4.19 A Relational response to the Diet and Cholesterol task on the post-assessment……119

4.20 A Relational response to the Speed Trap task on the post-assessment……121

4.21 A Prestructural response to the Class Scores task on the pre-assessment……...123

4.22 A Unistructural response to the Cholesterol and Diet task on the pre-assessment……124

4.23 A Unistructural response to the Migraine Treatment task on the pre-assessment……125

4.24 A Prestructural response to the Migraine Treatment task on the post-assessment……126

4.25 A Prestructural response to the Speed Trap task on the post-assessment……127

4.26 A Unistructural response to the Speed Trap task on the post-assessment……...131

4.27 A Multistructural response to the Speed Trap task on the post-assessment……132

4.28 A Relational response to the Speed Trap task on the post-assessment……133

4.29 A Prestructural response to the Speed Trap task on the post-assessment……134

4.30 A Prestructural response to the Hiring Discrimination task on the post-assessment……135

4.31 Levels of formal inferential reasoning responses provided on the post-assessment tasks by area of certification……137

4.32 Frequency of coded tasks by mathematical proficiency strands……144

4.33 Frequency of coded tasks by content area……...147

4.34 Frequency of coded tasks by content area without probability……149

4.35 Frequency of coded tasks by core concepts……151

4.36 Frequency of coded tasks by aggregate concepts……151

ABSTRACT

This study characterizes how a cohort of 33 middle and secondary mathematics preservice teachers' inferential reasoning changed while enrolled in a statistics course designed for future teachers. Changes in inferential reasoning from pre- to post-assessments are analyzed and further elucidated by midcourse clinical interviews conducted with a stratified random sample of 12 participants. Using a modified SOLO taxonomy (Biggs & Collis, 1982, 1989), the average dominant level of inferential reasoning for the cohort shifted from Unistructural to Multistructural over the course. However, considerable variation was evident at the cohort-level *within* specific tasks, and at the preservice teacher-level *across* tasks. While 58% of all participants increased their level of inferential reasoning, growth was more pronounced for secondary teachers with 75% increasing one or more levels compared with 50% for the middle school teacher population. A relationship between informal and formal approaches to inferential tasks was determined as 80% of levels assigned to formal inferential task responses were concordant with the dominant informal inferential reasoning level. Classification of 375 course tasks by *mathematical strands of proficiency* (Kilpatrick et al., 2001) revealed an increased demand for adaptive reasoning occurs simultaneously with the introduction of formal inferential methods. Prior to the topic of statistical inference, the primary proficiency strands emphasized by tasks are conceptual understanding (56%) and procedural fluency (75%). The concepts of center, variability and sample were heavily emphasized in the course while sampling variability was given little attention. Implications for research and the statistical preparation of teachers are offered.

CHAPTER 1: STATEMENT OF THE PROBLEM AND BACKGROUND

During the past quarter century, the ability to collect and analyze large quantities of data has been facilitated through rapid technology innovations. Statistical information, findings and claims are presented to the public regularly whether the context is choosing medical treatments, reading about economic trends in multimedia, watching an athletic event, or interpreting public opinion polls. People increasingly rely on statistical information and interpretations when making decisions as consumers, citizens and professionals. In response to these growing societal demands, statistics has become a key topic of the mathematics curriculum over the past 25 years (Franklin et al., 2007; Jones & Tarr, 2010).

The Purpose of the Study

Recently, national standards have placed a greater emphasis on statistics for middle and secondary school students (College Board, 2006; National Council of Teachers of Mathematics [NCTM], 1989; National Governors Association Center for Best Practices [NGA Center] & Council of Chief State School Officers [CCSSO], 2010). Until recently, many middle and secondary school mathematics teachers have not had an opportunity to learn statistics content during their college coursework. Therefore, many teachers are less prepared to teach statistics in comparison to other mathematics content strands (Conference Board of the Mathematical Sciences [CBMS], 2001). As a result, teachers have difficulties in both understanding and teaching the core ideas of statistics (Garfield & Ben-Zvi, 2008).

In a comprehensive review of the literature in statistics education, Shaughnessy (2007) argues there has been limited research regarding the preparation of middle and secondary mathematics teachers specific to statistics. Therefore, a need exists to study how well prepared middle and secondary preservice teachers are in terms of knowledge needed for teaching statistics. Currently, researchers and teacher educators claim little is known about the knowledge needed to effectively teach statistics in middle and secondary school settings. However, the ability to apply the statistical content and processes to be taught is clearly a minimum requirement. This study focuses on characterizing how a cohort of preservice middle and secondary mathematics teachers' statistical reasoning changes during a semester long statistics content course. More specifically, I focus on changes in preservice teachers' inferential reasoning as they progress through a statistics course designed exclusively for teachers.

In order to define inferential reasoning for purposes of this study, two broader concepts must also be described. First, *statistical inference* refers to moving beyond the data at hand in order to make decisions about some wider universe, taking into account that variation is everywhere and that conclusions are therefore uncertain (Moore, 2004). Second, *statistical reasoning* is defined "as the way people reason with statistical ideas and make sense of statistical information" (Garfield & Ben-Zvi, 2004, p. 7). Hence, *inferential reasoning* is the way that people make sense of statistical ideas and information in order to generate a conclusion that extends beyond the data at hand.

Given the wide application of inferential reasoning techniques, teaching students to generate and evaluate inferences based on realistic data has become a key goal of middle and high school statistics education (Franklin et al., 2007; NCTM, 2009; NGA

Center & CCSSO, 2010). During middle school, students are taught to coordinate statistical information informally to create inferences and make predictions. *Informal inferential reasoning* is accomplished by comparing two or more aspects of data when generating an inference without the assistance of a formal algorithm. In secondary school, students transition to *formal inferential reasoning* which necessitates the use of formulas, calculations and statistical tables to yield formal statistical inferences. In both approaches to inferential reasoning, students must attend to the context in which the data resides and the ever-present existence of variation.

The purpose of this study is to characterize how a cohort of middle and secondary preservice mathematics teachers' inferential reasoning changes while enrolled in a statistics content course. In order to provide a context for the changes, the cohort's opportunity to learn inferential reasoning is carefully documented through a task analysis and a description of both the statistical content taught and the emphasis placed on statistical reasoning. Specifically, the statistics content taught is classified into the following categories: measures of center, skewness, spread, variance, distribution, probability, sampling, variability, sampling variability, and inference. In addition, the emphasis placed on statistical reasoning is portrayed through a classification of tasks onto the five strands of mathematical proficiency: conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition (Kilpatrick, Swafford & Findell, 2001).

In summary, Shaughnessy (2007) notes that very little research has been conducted with preservice mathematics teachers who have had an opportunity to learn statistical content and processes during college coursework. From the few research

studies conducted in this vein, initial indications are that teachers still lack confidence in their statistical content knowledge and teaching of statistics despite an increased opportunity to learn (Groth & Bergner, 2005; Leavy, 2006). "The studies [that] focused on preservice and in-service K-12 teachers suggest that both have difficulties understanding and teaching core ideas of probability and statistics" (Garfield & Ben-Zvi, 2008, p. 28). If teacher educators are unable to develop the statistical reasoning of preservice teachers, then teachers certainly will be ill prepared to provide effective statistics instruction once inside the classroom. The present study provides a needed assessment of preservice teachers' ability to inferentially reason *before* and *after* participating in a statistics course designed specifically for them. The findings inform teacher educators and curriculum developers of middle and secondary mathematics teachers.

Research Questions

The study addresses the following three research questions:

1. How can the change in middle and secondary preservice teachers' inferential reasoning abilities be characterized during a statistics course?
2. Does a relationship between preservice teachers' change in informal and formal inferential reasoning exist? If so, how can it be characterized?
3. What opportunities to learn inferential reasoning are afforded middle and secondary preservice teachers during a semester-long statistics course?

Teachers' Statistical Reasoning

Given that statistics has only recently been considered a core content area in the K-12 curriculum, most teacher preparation programs have historically dedicated scant

attention to this domain. In 2001, Watson developed a tool to assess teachers' statistical content knowledge, pedagogical content knowledge, knowledge of teaching and confidence for teaching statistics. Teachers, ranging from kindergarten to grade 10 expertise, reported that they lacked confidence in their knowledge of statistics and also that they received few if any opportunities for professional development related to data and chance. According to the College Board of Mathematical Sciences (2001), "Of all the mathematical topics now appearing in the middle grade curricula, teachers are least prepared to teach statistics and probability" (p. 114). In response to this situation, a number of programs and materials have been created to address professional development needs of mathematics teachers. Recently, several studies assessed how teachers' statistical reasoning evolved during the course of professional development (Heaton & Mickelson, 2002; Makar & Confrey, 2004; Rubin & Rosebery, 1998). Shaughnessy (2007) states:

> Most K-12 mathematics teachers in the United States have very little background in statistics. The exceptions are those teachers who may have had a concentration in statistics during their masters program for secondary teachers, or middle school teachers who completed one of the few special programs that exist in the United States for middle school mathematics teachers. (p. 995)

While work is clearly underway to provide opportunities for preservice and inservice teachers to increase their statistical content and pedagogical knowledge, the learning goals associated with statistics education are shifting from learning formal procedures toward reasoning and sense making. According to Moore (1997), "[A] grasp of the reasoning of inference is more important than how many individual procedures"

are learned in a given statistics course (p. 127). Based on findings from research with tertiary students, Garfield and Ben-Zvi (2007) report that students who successfully complete statistics courses are able to apply inferential statistical methods, but often do not understand why the methods are appropriate. In addition, these students also lack a solid understanding of core statistical concepts such as measures of center, variation and distribution. From the few research studies involving K-12 teachers, teachers have difficulty generating inferences and applying formal statistical knowledge to real-world scenarios (Heid, Perkinson, Peters & Fratto, 2005; Liu & Thompson, 2005). Therefore, teachers' statistical reasoning is often inconsistent from test-item to test-item and from topic-to-topic. Garfield and Ben-Zvi (2007) attribute a portion of the problem to the historically procedural approach to teaching statistics but emphasize the difficulty and complexity of developing reasoning skills.

Inferential Reasoning

Inferential reasoning has served as a unifying theme of introductory statistics courses at the tertiary level for a number of years (Konold & Pollatsek, 2002). With the recent emphasis of statistical reasoning in middle and secondary schooling, the unifying role of inferential reasoning is gaining in prominence (NGA Center & CCSSO, 2010). Current recommendations for middle and secondary statistics education outlined in the *Guidelines for Assessment and Instruction in Statistics Education [GAISE]* report support the introduction of inferential reasoning during middle school informally and then formalization of inferential reasoning in secondary years (Franklin et al., 2007). Generally, two types of problems fall under the broad definition of inferential reasoning: (a) generalizing from samples to populations, and (b) comparison and determination of

cause from randomized comparative experiments (Garfield & Ben-Zvi, 2008). These two problem types can employ formal statistical methods or be accomplished through informal approaches. Informal approaches allow students to engage in inferential learning at an earlier age, and studies have shown that upper-elementary age students can successfully draw inferences (Stohl & Tarr, 2002; Watson, 2002; Watson & Moritz, 1999).

Developing Inferential Reasoning Abilities

Researchers believe that introducing inferential reasoning informally assists students in developing argumentation structures necessary for understanding formal methods (Wild & Pfannkuch, 1999). More generally, researchers have found that students who practice informal reasoning throughout schooling develop rich mental schemas that aid in problem solving and also enhance future learning (Means & Voss, 1996). Means and Voss propose introducing argumentation early in schooling to scaffold students' learning. Informal reasoning is a global skill that increases learning across content domains. Through the process of informal reasoning, students construct *situation models* in their minds that help them connect knowledge in meaningful ways and generate inferences. Situation models are especially relevant to the domain of statistics as context plays a critical role in defining the problem to be solved and how to interpret results. Students who create situation models will frame the problem statement using the context of the task and connect relevant pieces of information together in a logical manner to formulate an inference. Students who lack the ability or experience in developing situation models are less likely to develop robust arguments and may only be able to retrieve bits of disconnected facts. On the other hand, students with more highly

developed situational models are better than peers at learning, generating inferences, and problem solving.

Reasoning Summary

The importance of developing students' reasoning abilities has been acknowledged with the recent secondary mathematics recommendations endorsed by NCTM (2009). In statistics education, inferential reasoning plays a crucial role in uncovering how well students understand core statistical concepts, develop relationships between these concepts, and draw inferences based on data analyses. Students can reason inferentially at a young age, and current standards endorse reasoning inferentially with informal methods during middle school and progressing to formal strategies in secondary schooling (Franklin et al., 2007). The developmental model reflected in these recommendations aligns well with both researchers' conjecture that informal methods are needed to develop argumentation structures to support formal methods and findings from educational psychology that emphasize the critical role argumentation plays in enhancing future learning and drawing inferences (Means & Voss, 1996; Wild & Pfannkuch, 1999).

Conceptual Framework

The conceptual framework for this study examines the key knowledge components and processes needed to reason inferentially. The conceptual framework consists of two developmental stages. As depicted in Figure 1.1, Stage 1 relates to the development of informal inferential reasoning, while Stage 2 relates to formal inferential reasoning. The stages of development imply that preservice teachers are able to both apply informal and formal methods to generate inferences and explain why the procedures are appropriate.

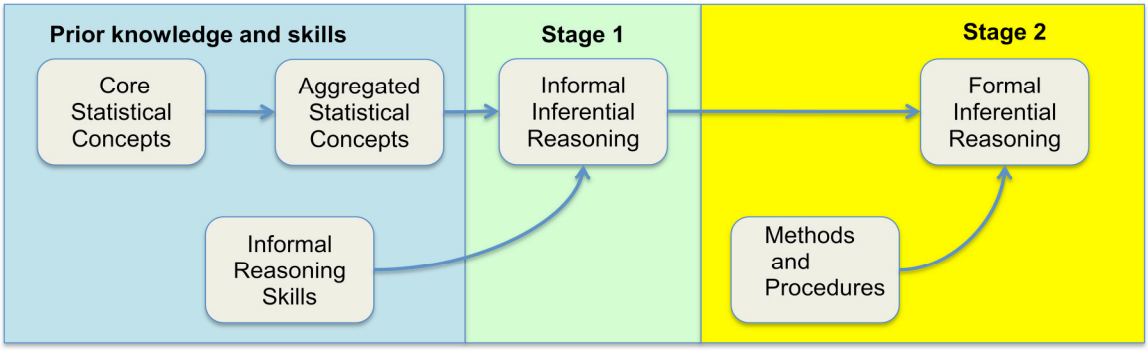


Figure 1.1. Conceptual framework for inferential reasoning development

Stage 1: Informal Inferential Reasoning Components

Statistics educators advocate that a focus on developing informal inferential reasoning skills precede instruction of formal methods for inference. Informal inference is theorized as a bridge between exploratory data analysis (an unstructured activity) and formal methods that consist of prescribed steps and conditions for application (Wild & Pfannkuch, 1999). In order to engage in productive informal approaches to statistical inference, students need to have experiences in argumentation or informal reasoning (Means & Voss, 1996), adequately developed core statistical conceptions, and adequately developed aggregate statistical conceptions.

Core statistical concepts are measures of center, spread, density, skewness and outliers (Reading & Read, 2006). Ideally, core statistical concepts are introduced and explored through the overarching idea of data distribution in the form of dot plots. Well-developed notions of core concepts and how they relate are needed to describe data distributions (Bakker, 2004). Once this foundation has been developed, *aggregate statistical concepts* are introduced and consist of distribution, sample, variability, and sampling variability (Zieffler, delMas, Garfield & Gould, 2007). These aggregate

conceptions are essential for inferential reasoning, and are commonly introduced and explored in the beginning phases of generating an inference. Typically, a three-step transition occurs from first comparing two complete populations using core statistical concepts, then inferring population characteristics based on sample data utilizing sampling techniques and accounting for sampling variation, and finally creating inferences about two unknown populations from random samples by way of formal hypothesis tests or confidence intervals.

Unfortunately, core and aggregate statistical concepts are often taught only through the use of procedures, which hinders conceptual understanding (Garfield, delMas & Chance, 2007). Preservice teachers who possess only a procedural understanding of statistics concepts are not able to apply knowledge appropriately or populate their argumentation schema to generate a prediction that extends beyond the data at hand. Similarly, preservice teachers who do not have experience with informal reasoning may have difficulty in coordinating the demands of the problem statement with existing knowledge to produce a logical argument. Evidence of such a case would consist of a preservice teacher who demonstrates conceptual understanding of core and aggregate concepts but is unable to provide a logical prediction to an inferential task (Means & Voss, 1996).

Stage 2: Formal Inferential Reasoning Components

Once preservice teachers are able to demonstrate the ability to inferentially reason with informal methods, ideally their learning progresses to the next stage of development with a goal of productive formal inferential reasoning. The additional component required for generating formal inferences beyond those of informal approaches is a sound

understanding of formal methods and procedures. Typically, the second half of an introductory to statistics course commonly entails statistical methods and procedures to support either generalizing from samples to a larger universe or drawing inferences related to the comparison of two data sets. At this stage, preservice teachers' require the addition not only of knowledge of formal statistical methods but also an understanding of when to use the methods, why they work, and what the results do and do not mean.

In summary, the conceptual framework documents a compilation of current statistic educators' and researchers' recommendations for learning how to reason inferentially. The progression begins with the development of core statistical conceptions and terminates with the selection and application of formal inferential methods. The goal of the progression is to ensure that preservice teachers are able to inferentially reason with formal methods in a meaningful way and answer questions about why the processes are effective, how they work, and what the results do and do not imply. In addition, the preservice teachers should also be able to informally reason and draw inferences in situations where the specificity of formal reasoning is unnecessary.

Theoretical Framework

The selection of a theoretical framework is critical because it informs the research design, including data collection, data analysis, and interpretation of findings. In this study, I have selected a cognitive framework to characterize the developmental stages of statistical reasoning.

The Structure of Observed Learning Outcomes (SOLO) Taxonomy

The development of statistical reasoning has been characterized from cognitive perspectives in the past, and general agreement exists that students' learning progresses

through a number of hierarchical levels and cycles (e.g. Jones, Langrall, Mooney & Thornton, 2004; Mooney, 2002; Watson, Collis, Callingham, & Moritz, 1995). Cognitive models of development have evolved from maturation-only perspectives to models that account for both maturation and interactions experienced by the learner. In addition, neo-Piagetian, cognitive development theorists have refined stage-theory models of learning to characterize domain-specific learning rather than Piaget's universal stage model (Jones et al.). The most widely used cognitive model of students' development of statistical reasoning is the Structure of Observed Learning Outcomes (SOLO) taxonomy developed by Biggs and Collis in 1982. Refinements to the initial taxonomy acknowledge the existence and importance of multimodal learning, which positions earlier learning as foundational to later learning rather being replaced. The developmental model also acknowledges that learning can occur in a *top down* manner in addition to the typical *bottom up* approach as shown in Figure 1.2.

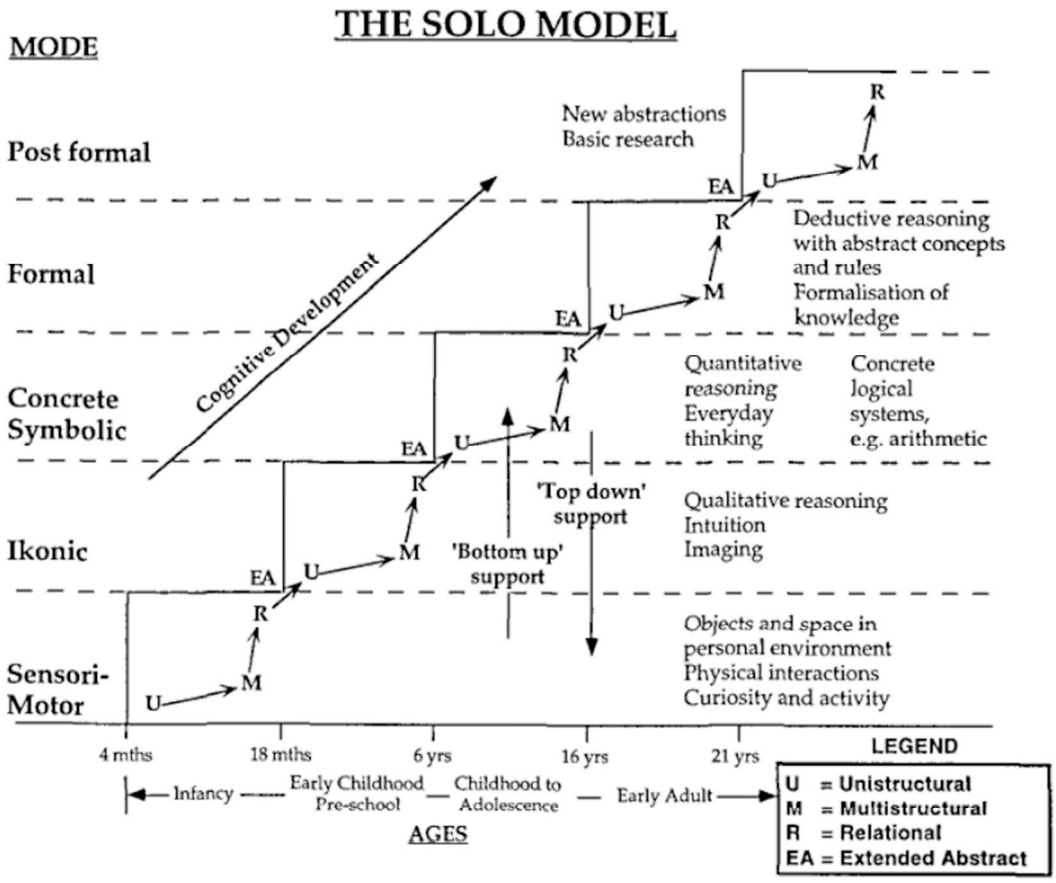


Figure 1.2 *The SOLO model*. From "A Model for Assessing Higher Order Thinking in Statistics," by J. M. Watson, K. F. Collis, R. A. Callingham and J. B. Moritz, 1995, *Educational Research and Evaluation, 1*(3), p. 249. Reprinted with permission.

According to Biggs and Collis (1982), within each mode or stage of learning, a cycle of five hierarchical levels of learning exists: (a) Prestructural, (b) Unistructural, (c) Multistructural, (d) Relational, and (e) Extended Abstract. The first level, *Prestructural* (P), marks the last level of development in the prior mode or stage of learning. Similarly, the final level, *Extended Abstract* (EA) is thought to be the first level at the next stage of development. Hence, the middle three levels characterize the growth of students' reasoning across a stage of development and will be the main focus for this research study. A cycle within each stage is called a U-M-R cycle to correspond with the first

letter in each level. A *Unistructural* (U) response to a task uses one relevant aspect of the mode. A *Multistructural* (M) response incorporates several relevant aspects of the mode in a disjointed and often sequential manner. Finally, a *Relational* (R) response is denoted by the productive coordination of relevant aspects and demonstrates an integrated understanding. Perturbations assist in advancing learning within a U-M-R cycle and also to new modes of reasoning. Without schooling or another method for challenging thinking, learning will generally remain at the concrete symbolic mode.

Characterization of Informal Inferential Reasoning Responses

In a pivotal study with middle school students, Mooney (2002) reported that students provide P, U, M and R responses to informal inferential reasoning tasks corresponding in the concrete symbolic mode. In addition to identifying these four levels of responses, Mooney described three underlying subprocesses associated with students' informal inferential reasoning: (a) making comparisons *within* data sets or data displays, (b) making comparisons *between* data sets or data displays, and (c) making inferences *from* a given data set or data display. By describing these three subprocesses in relation to the P, U, M and R levels of the SOLO taxonomy, Mooney augmented the cognitive framework with descriptions of informal inferential reasoning at the first four levels. In this study, changes in preservice teachers' inferential reasoning are assessed at three time points during the semester. The characterization of the cohort's informal reasoning responses is completed utilizing the original SOLO taxonomy augmented by Mooney's informal inferential reasoning descriptors.

Characterization of Formal Inferential Reasoning Responses

The SOLO taxonomy identifies a formal mode of learning immediately following the concrete symbolic mode (Biggs & Collis, 1982). However, no research studies have utilized the SOLO taxonomy for characterizing formal inferential reasoning responses in a manner similar to the research conducted by Mooney in the concrete symbolic mode (2002). Therefore, the cohort's formal responses to inferential reasoning tasks are coded to the general categories of P, U, M and R based on the original definitions developed by Biggs and Collis (1989).

Response Classification

In addition to categorizing each response to assessment tasks as either P, U, M or R, the responses are also classified by core and aggregate statistical concepts. For example, if a response consists of an argument comparing the means of two data displays, then the response is classified as focused on the concept of *center*. If the response makes reference to one display that has data "spread out" while the other display has data "clustered" together, *variance* is assigned as well. The purpose of this additional layer of coding is to understand what concepts participants were drawing upon during informal inferential reasoning throughout the semester.

Reasoning Progression

Specific to the topic of inferential reasoning, Reading (2007) was the first to recommend that students' cognitive development be characterized using a two-stage SOLO taxonomy with informal reasoning at the first stage and formal at the second. Rather than assume this progression and dependency between informal and formal inferential reasoning, the cohort's responses are characterized by a one-cycle SOLO

taxonomy in the concrete symbolic learning mode for informal responses and by a one-cycle SOLO taxonomy in the formal mode of learning for formal responses. In order to justify or refute a two-cycle progression, the existence of a relationship between informal and formal changes in reasoning was sought once all responses had been characterized.

Significance of the Study

The study contributes to the statistics education community in several ways. First, the cohort of preservice teachers in the study experienced an ideal situation in terms of tertiary opportunities to learn statistics. More specifically, the cohort participated in a statistics content course that was designed specifically for middle and secondary preservice teachers rather than an introductory statistics course for a general student population. In this manner, the study analyzes the learning of preservice teachers in a potentially ideal context. As stated previously, research on the development of preservice teachers' statistical content knowledge is lacking. After reviewing research literature, Garfield and Ben-Zvi (2008) state that, "The studies suggest further explorations are needed in the issues of developing teacher knowledge of statistics as well as methods of helping teachers to understand the big ideas of statistics" (p. 28). Therefore, the findings in terms of these preservice teachers' statistical reasoning abilities serve as a barometer for how adequately or inadequately prepared future teachers will be to fulfill the recommendations put for in the *GAISE* recommendations (Franklin et al., 2007).

Secondly, the study provides further evidence supporting or refuting the need for informal inferential reasoning as a necessary predecessor to formal inferential reasoning. While a causal link between the two learning milestones is beyond the scope of this study, the relationship between the two types of reasoning is examined in detail and

provides guidance to future statistics education efforts. Prior research studies geared toward developing inferential statistical reasoning produced disappointing results with students unable to articulate how to generate inferences or conduct hypothesis tests (Zieffler et al., 2007). In addition, college students have been unable to consistently describe core and aggregate statistical concepts (Garfield, delMas & Chance, 2007). Statistics education researchers advocate that informal inferential reasoning serves as a needed cognitive milestone between exploratory data analysis and formal inferential methods (Zieffler et al., 2007). Accordingly, some statistics educators are beginning to revise introductory statistics courses to include informal inferential reasoning. However, an examination of a student's ability for informal inferential reasoning and then the corresponding capacity for formal inferential reasoning has yet to be conducted (Garfield & Ben-Zvi, 2008).

More importantly, the findings of this study assist instructional efforts for both middle and secondary students and preservice teachers. As Shaughnessy (2007) notes, students and teachers often experience the same challenges when learning statistics. The cognitive models of change in statistics are especially valuable because they provide needed guidance about student learning that can inform curricular design, sequencing of topics, task creation, and assessment design. "Because these models incorporate domain-specific knowledge of students' statistical reasoning across key statistical concepts and processes, they arm teachers with the kind of knowledge that can be used in the design, implementation, and assessment of instruction in statistics and data exploration" (Jones et al., 2004, p.112). Specific to use of the SOLO framework, Shaughnessy states, "The SOLO model has been genuinely useful in helping to describe student reasoning on a

number of concepts in statistics like average, variability, comparison of data sets, and so on" (p. 1001). In addition, the SOLO framework supports multiple modes of thinking and levels of progression within each mode. This structure provides versatility in characterizing thinking and reasoning at a variety of levels and accommodates multiple approaches to tasks.

Summary

This study characterizes how a cohort of middle and secondary preservice teachers' inferential reasoning changes during a statistical content course. Leaders in statistics education and educational psychologists advocate the need to introduce statistical inference through informal methods first in order for formal approaches to make sense and be understood. Thus, a SOLO framework consisting of two learning modes, informal and formal, is employed to characterize the change of the preservice teachers' inferential reasoning. The first mode focuses on the development of core and aggregate concepts and processes related to inferential reasoning using informal approaches. The second mode again requires an understanding of core and aggregate statistical concepts but utilizes formal approaches to generate inferences. The findings of this study establish a needed characterization of preservice teacher knowledge specific to inferential reasoning consistent with the goals of the *GAISE* report (Franklin et al., 2007). In addition, the importance of developing informal approaches as a foundation for the learning of formal methods is explored. Finally, the opportunity to learn inferential reasoning afforded by this statistics course specifically designed for preservice mathematics teachers is carefully analyzed and provides a rich description of the learning context.

The following four chapters are organized to provide a comprehensive account of the research study. First, I provide a review of relevant literature in chapter 2, identifying the need and importance of addressing the research questions posed. Next, I describe the data collection processes and analysis methods in order to create transparency and enhance the validity of results in chapter 3. Then, I provide detailed research findings in chapter 4. Lastly, I discuss the results and offer implications for teacher education and future research in chapter 5.

CHAPTER 2: LITERATURE REVIEW

The conceptual framework outlined in the previous chapter underscores the complexity of reasoning about statistical inference and the usefulness of addressing the research questions. In this chapter, a detailed review of research related to reasoning about statistical inference is presented. The literature review is organized into five sections: (1) the rise of inferential reasoning in the mathematics curriculum, (2) students' understandings of inferential reasoning, (3) teachers' understandings of inferential reasoning, (4) characteristics of tasks used to assess inferential reasoning, and (5) a summary of key findings that inform the research design of this study.

The Rise of Inferential Reasoning in the Mathematics Curriculum

The Evolution of Statistics Standards

The history of statistics as a content strand in the school mathematics curriculum has differed from other content areas such as algebra and geometry. The first recommendations to include statistics in the school mathematics curriculum occurred in a report named *The Reorganization of Mathematics in Secondary Education* in 1923 and advocated that middle school students learn how to create and interpret of graphs, and that secondary students be offered the opportunity to learn about measures of central tendency as an optional topic (Tarr & Jones, 2010). The progression of recommendations regarding the inclusion of statistics as a content strand for all grades in the school mathematics curriculum is shown in Figure 2.1.

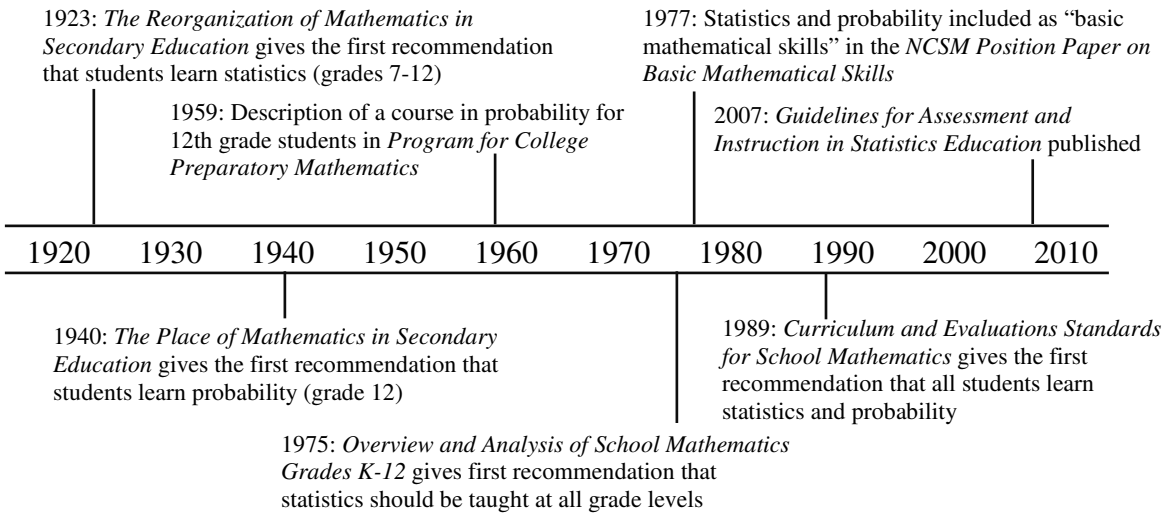


Figure 2.1. Timeline of shifting emphases on statistics and probability in the K-12 school curriculum. From "Recommendations for Statistics and Probability in School Mathematics Over the Past Century," by J. E. Tarr and D. L. Jones, 2010, In B. J. Reys, R. E. Reys, and R. Rubenstein (Eds.), *Mathematics Curriculum: Issues, Trends, and Future Direction: Seventy-second Yearbook* (pp. 65-76). Reston, VA: National Council of Teachers of Mathematics. Reprinted with permission.

The evolution of statistics in the school mathematics curriculum broadly occurred across three eras (Tarr & Jones, 2010). The first era, from 1923 to 1959, emphasized creating and interpreting data displays primarily in the form of tables and graphs. Measures of central tendency were viewed as advanced topics and targeted for only the most mathematically able students in secondary school. Hence, statistical content was relatively limited. During the second era of 1960 through 1979, statistical content previously reserved for secondary school was now recommended for either late elementary or middle school grades. In addition, the view that statistics content was only for the most mathematically able students was altered in the 1970s, allowing all students to partake in learning statistics during middle school years. More advanced statistics topics emerged as well and were offered during secondary school years. The final era spans the 1980s through current times with statistics emerging as a prominent content

strand in the school mathematic curriculum across grades K through 12. Coupled with a focus on problem solving and the availability of technology tools, statistics content is often taught by answering real-world questions with the use of calculators or other technology tools.

In the 1970s and 1980s, advancements in technology, data availability, and more sophisticated statistical approaches created momentum to alter the type of statistics content available in K-12 curricular materials. In 1984, the *Quantitative Literacy Project* was funded by the National Science Foundation (ASA-NCTM, 1984). The project was initially launched by the Joint Committee on the Curriculum in Statistics and Probability, which was comprised of members from the American Statistical Society and the National Council of Teachers of Mathematics. The members of the *Quantitative Literacy Project* attempted to accelerate integration of up-to-date statistical content in the K-12 mathematics curriculum through new curricular materials and teacher professional development activities. Four books were developed by a team of statisticians and teachers, *Exploring Data, Exploring Probability, The Art and Techniques of Simulation,* and *Exploring Surveys – Information from Samples.* This curriculum development project focused on investigative approaches to learning statistics rather than application of formal methods and algorithms.

In 1989, the National Council of Teachers of Mathematics (NCTM) released the *Curriculum and Evaluation Standards for School Mathematics*, which included statistics and probability among the mathematics content strands for the K-12 curriculum. The authors of this pivotal document stated that statistics and probability are used to make decisions about marketing, defense, business and research. Hence, all students should

have the opportunity to learn probability and statistics content to increase their career opportunities and to become informed citizens. In 2000, NCTM updated content and process recommendations in *Principles and Standards for School Mathematics,* and again emphasized the importance of data analysis and probability for all students. In addition to NCTM's endorsements, in 2006 the College Board released, *College Board Standards for College Success: Mathematics and Statistics,* advocating that data, variation, chance, fairness and risk be included as core content for middle and secondary mathematics students.

Despite strong commitments and guidance from NCTM and the College Board for greater emphasis on statistical content in the K-12 curriculum, more detailed recommendations were needed. In an effort to support classroom teachers and inform the development of curricular materials, a joint committee of the American Statistical Association and NCTM funded the development of the *GAISE* recommendations (Franklin et al., 2007). Specifically, the *GAISE* recommendations complement NCTM's Data Analysis and Probability standards of *Principles and Standards for School Mathematics* by providing a framework for statistical problem solving and an articulation of the progression of statistical concepts and processes for preK-12 students.

Franklin et al. (2007) realized that the key to implementing statistics content standards resides with not only providing a detailed curricular framework, but also with providing educational support for teachers. Researchers have observed that although standards place significance on teaching statistics, changes in actual classrooms lagged recommendations (Jones et al., 2004). One possible explanation proposed is that teachers have not experienced learning statistics content and processes in alignment with current

standards. Franklin et al. (2007) state, "Statistics […] is a relatively new subject for many teachers, who have not had an opportunity to develop sound knowledge of the principles and concepts underlying the practices of data analysis that they are now called upon to teach" (p. 5). Hence, many teachers do not understand how statistics differs from traditional mathematics or how to organize statistical topics into a coherent and cohesive curricular strand.

Endorsing Reasoning and Sense-Making

Specific to secondary school, NCTM recently released *Focus in High School Mathematics: Reasoning and Sense Making,* a series of professional development materials. A key principle of these documents is to shift focus away from covering specific topics in high school mathematics to unifying teaching, curriculum and research efforts around the development of students' reasoning and sense-making abilities. With enhanced reasoning abilities, it is expected that students will be able to solve real-world problems in a variety of contexts, compete in a global workplace, succeed in future learning and become productive citizens (NCTM, 2009). A similar shift in philosophy and priorities has emerged in statistics education as well, as NCTM (2009) argues, "The development of statistical reasoning must be a high priority for school mathematics" (viii).

In addition to changes in the composition of curriculum standards and the development of curriculum materials, the focus of statistics education research has likewise changed in recent years. According to Garfield and Ben-Zvi (2004), past [research] efforts primarily entailed improving instruction and learning related to statistical techniques, formulas, and procedures. However, more recent studies

emphasize the development of students' statistical reasoning and thinking as the primary educational goal. According to Garfield and Ben-Zvi (2008):

> The topics of these research studies conducted by members of this community [the International Statistical Reasoning, Thinking and Literacy Research Forums] reflect the shift in emphasis in statistics instruction, from developing procedural understanding, i.e., statistical techniques, formulas, computations and procedures, to developing conceptual understanding and statistical literacy, reasoning, and thinking. (p. 35)

The recent change in emphasizing reasoning and sense-making coupled with a heavier emphasis on statistics content in the middle and secondary grades creates a gap between teacher experiences as learners and what they will teach in the future.

Inferential Reasoning

Recently, the NGA Center and CCSSO placed explicit attention on the development of inferential reasoning in the mathematics school curriculum by requiring that middle school students learn how to informally reason about inference and secondary students learn how to formally reason about inference in the Common Core State Standards (2010). As more states pledge adoption, the Common Core State Standards for Mathematics are anticipated to become the *de facto* national mathematics school curriculum. Since preservice teachers' learning of statistics most likely did not progress from informal to formal approaches, their conceptual knowledge may be lacking to support these new standards. Informal approaches to generating inferences are relatively new, and require the coordination of core and aggregate statistical concepts (Figure 1.1). Therefore, knowledge of these concepts and how they are related is required. Ideally,

these concepts are also developed and understood when learning about formal approaches to inferences. However, formal inference has traditionally been taught in a procedural manner without a focus on conceptual understanding or reasoning (Garfield & Ben-Zvi, 2007). Hence, statistics education as experienced by preservice teachers may have very different connotations than the learning expectations supported by the Common Core State Standards and similar standards documents.

The Process of Inferential Reasoning

Because so few research studies have been conducted specifically with preservice and inservice teachers, most of what is known about the process of learning how to inferentially reason is based upon student populations. Therefore, the studies and findings presented in this section relate to research studies with students. A synthesis of the few research study findings specific to teachers will follow the inferential reasoning summary.

Inferential reasoning is a complex and challenging endeavor. As mentioned previously, two types of inferential tasks exist: (a) generalizing from samples, and (b) comparison and determination of cause from randomized comparative experiments (Garfield & Ben-Zvi, 2008). Historically, statistical inference has been relegated as a topic for college-level statistics courses. In such courses, students are taught how to generate inferences by means of formal approaches and procedures, such as hypothesis testing during college coursework. Such approaches have failed to yield sound statistical reasoning capabilities. In particular, numerous studies indicate that upon completion of introductory college statistics courses, many students: (a) continued to struggle with every aspect of formal approaches of generating inferences (Aquilonius, 2005), (b) could

not explain basic statistical concepts (Garfield et al., 2007), and (c) were unable to identify the relationships between statistical concepts (Williams, 1999). In addition, researchers have not found a relationship between college students' academic performance in statistics courses and their understanding of statistics (Tempelaar, Van der Loeff & Gijselaers, 2007). Hence, a reform movement is currently underway to improve how statistics is taught at the college level, and researchers and college instructors are exploring alternative instruction methods that encourage a rich, integrated understanding of core statistical concepts and processes (Garfield et al., 2007).

Data Distributions

Researchers have sought to identify the origins of inferential reasoning. In a synthesis of research studies, Reading (2007) identified a set of statistical concepts (variation, distribution, mean, spread, and graphs) and actions (focus on proportions, sample variation, and randomness) that needed to be understood prior to engaging in informal inferential reasoning. These concepts and actions align well with the conceptual framework provided in Figure 1.1. Reading's analysis supports that well-developed core statistical concepts are a prerequisite to engaging in inferential reasoning. In addition, an understanding of one aggregate concept, data distribution, is also a prerequisite. The following section discusses the unique role of data distribution in connecting and relating core statistical concepts in the early stages of inferential reasoning.

The authors of the *GAISE* recommendations introduce statistical inference through the comparison of two complete populations that are represented graphically as data distributions; in doing so, students are well positioned to compare important characteristics of the data sets such as center, spread and shape visually (Franklin et al.,

2007). Data distributions build upon students' knowledge of core statistical concepts and serve as a unifying structure for exploring the relationships between them. Although distribution is an aggregate concept that normally is not discussed by name until middle or secondary schooling, students are afforded experiences in creating data displays (i.e., graphical distributions), typically in elementary school years (Watson, 2009). By describing the range, spread and shape of data distributions, elementary students begin to develop a sense of variation and a notion of center. Therefore, data distributions foster a shift in students' views from single data points or outcomes to global characteristics.

Given that relatively few students in middle and secondary school have been exposed to the ways of thinking needed for data analysis and interpretation, researchers have explored various ways to build inferential thinking while also developing necessary conceptions such as center, spread, shape and sample. Most researchers have focused on middle school students because they have been taught how to create displays of data distributions (e.g., bar charts, histograms) and calculate measures of center such as mean and median. In order to informally generate inferences about data sets, students must attend to global characteristics and trends, such as the shape of a data distribution or how data is condensed about an interval (Cobb, McClain & Gravemeier, 2003). However, because core statistical concepts such as measures of center are often introduced in a rigid and procedural manner, this can impede a student's ability to view distributions globally and therefore, impair inferential reasoning. Cobb et al. (2003) found that the operational view of median held by many middle school students impedes their ability to coordinate the concept of center with other aspects of a data distribution and inferentially reason. However, Bakker (2004) reported success in developing the concept of distribution with

middle school students as a result of a guided reinvention classroom design experiment. Students contributed their own ideas, strategies, and language when solving statistical problems and defined data distribution and other core statistical concepts together as a class. The instructor, Bakker, led the process to ensure that the activities and discussions reinvented or aligned with accepted statistics practices. Through this sense-making process, Bakker was able to foster students' understanding of core concepts from their intuitive notions by applying these ideas to more complex scenarios involving data distributions.

In a research study with students in grade 3, 5, 6, 7, and 9, Watson and Moritz (1999) found that students who employ averaging strategies, including calculating the arithmetic mean, visualizing the mean, or finding a balance point, are more successful in comparing two populations of unequal size than their peers. In addition, Watson and Moritz determined that students who do not proportionally reason are typically unsuccessful in comparing data sets of unequal size. Therefore, in order to be successful in this first step towards informal inference, comparison of two populations, students must possess: the ability to reason proportionally, a basic understanding of core statistical concepts, and knowledge of how they relate to form a data distribution. However, other researchers dispute the notion that proportional reasoning is a prerequisite to inferential reasoning. For example, Stohl and Tarr (2002) found that grade 6 students were able to generate inferences through the use of technology tools and multiple views of data representations.

Sample and Sampling Variability

Two other key aggregate concepts (Figure 1.1) required for inferential reasoning are sample and sample variability. While students may generally possess experiential knowledge related to distribution from elementary school, often sampling is a topic relatively unexplored prior to middle school. The idea of selecting a random sample from a population necessitates the need to understand probability distributions and the associated likelihood of selecting various elements of the data set in relation to that distribution. Similar to the notion of distribution, students often struggle in viewing a sample globally, but rather focus on each individual data point. Saldanha and Thompson (2003) studied the development of secondary "students' thinking as they participate[d] in instruction designed to support conceiving sampling as a scheme of interrelated ideas including repeated random selection, variability among sample statistics, and distribution" (p. 259). Saldanha and Thompson determined that students tend to focus on individual samples and statistical summaries of samples instead of how collections of samples are distributed.

Similarly, Pratt, Johnston-Wilder, Ainley and Mason (2008) studied how students shift attention from immediate sampling results at the local level to global trends or aggregations of multiple sampling trials. Unfortunately, most students tended to focus on *local* changes between samples, which resulted in small shifts in aggregated results, rather than seeking global trends in the data. Pratt et al. hypothesized that student expectations of invariance were never fully met, causing students to distrust results. Consequently, they advocate the need to discuss the concept of sampling variation from the onset of instruction to assist students in global comparisons versus attending to

changes in local characteristics. In a related study involving sampling with the use of simulations, Stohl and Tarr (2002) found that students were able to effectively generate informal inferences with the assistance of multiple representations. Therefore, the representations and tools associated with the sampling process can impact students' ability to attend to global characteristics and generate inferences.

Research in cognitive psychology determined that intuitive or naïve misconceptions regarding the probabilistic elements of sampling can grow over time if left unattended. Considered one of the classics in statistics education, Tversky and Kahneman (1982) identified a pervasive network of misconceptions that they refer to as "the law of small numbers." The law of small numbers is similar to the law of large numbers, but also contains an extra element of self-correction, which occurs in steady-state systems but not in random processes. The law of small numbers is composed of two aspects: (a) the belief that small samples randomly selected from a population will embody characteristics of the population to such a degree that the sample is highly representative of the overall population, and (b) the gambler's fallacy, which assumes that laws of chance include an element of self-correction or fairness. For example, if a sequence of events has strayed away from a theoretically expected result, future outcomes will self-correct this discrepancy. In terms of sampling, the law of large numbers supports the notion that samples can be highly representative of the overall population, especially when the sample selected is large in comparison to the entire group.

Fischbein and Schnarch (1987) determined that the belief in the law of small numbers solidifies and grows for many students over time, and therefore, recommend

intervention by middle school years. The researchers argue that if a student's general schema is adequate to reason about a particular task, then misconceptions diminish over time. However, if a student's general schema is inadequate to deal with the constraints of a particular task, then the frequency of misconceptions increases as the student ages. In order to introduce cognitive conflict, researchers have employed the use of student-student interactions, prerecorded student comments, the careful sequencing of tasks, and the use of computer simulations to test and refute student conjectures about the law of small numbers. Collectively, these strategies for fostering more sound reasoning about sampling have achieved varying degrees of success.

Research indicates that working with middle school students through the use of simulation tools and cognitive conflict is successful (Stohl & Tarr, 2002; Watson, 2002), whereas researchers working with high school students in similar interventions are less successful (Pratt et al., 2008; Saldanha & Thompson, 2002). In addition, Saldanha and Thompson found that many students confused the *number of samples* with the *number of people* in the population. This misconception caused students to believe that sampling distributions were equivalent to the actual population. Lastly, Pratt et al. found that many [high school] students believed large samples produced unreliable results. One student voiced his concern with large sample sizes: "Because the 280 was just getting too stupid, I think, and had too much in" (pp.187-188). Other students felt that sample sizes of 100 were sufficient, since they perceived 100 to be a large number.

Variance

The remaining aggregate statistical concept (Figure 1.1) yet to be discussed explicitly is variance. Variation is the hallmark of statistics. According to Cobb and

Moore (1997), "the need for [statistics] arises from the *omnipresence of variability*" (p. 801). In a synthesis of research findings, Shaughnessy states that variation occurs in eight different forms: (i) variation in particular values such as outliers, (ii) variation over time, (iii) variation over an entire range, (iv) variation within a likely range, (v) variation from a fixed value such as a mean, (vi) variation in sums of residuals, (vii) variation in co-variation or association, and (viii) variation as a distribution (2007). Many of these types of variation can be observed graphically through data distributions and are introduced through the comparison of two samples or occur during the sampling process. Therefore, to some extent variance is embedded in aggregate statistical concepts. Watson (2008) found that students' perception of "reasonable" variation is closely tied to their knowledge and experiences with the context of a given statistical task. According to Watson, the more *relevant* the task context is to the student, the more *reasonable* the expectations for variation. Similarly, Pratt et al. (2008) found that students' tolerance for variance was much lower than expected due to their lack of experience in comparing empirical results to theoretical probability distributions. These studies suggest that students require multiple experiences embedded in different contexts with chance events in order to develop of sense of how much variance to expect.

Inferential Reasoning Summary

In summary, generating inferences requires students to possess robust statistical conceptions, to coordinate these conceptions, and to formulate a data-based argument situated in the context of the task. The process of inferential reasoning is a complex activity and at times counter-intuitive due to students' natural tendencies to possess naïve conceptions of probabilistic processes and focus on local phenomena. Therefore,

students need experiences throughout their schooling to develop an understanding of core statistical concepts, challenge intuitive misconceptions, explore relationships between concepts, experience chance events, develop inferences, and justify their arguments. Without such opportunities in middle and secondary schooling, research suggests that efforts to teach inference meaningfully at the college level are likely to be unsuccessful. While the middle and high school experiences of the cohort who participated in this study are not reported, they did experience a college level course designed to specifically meet their needs as future teachers who will be play a fundamental role in the implementation of both the Common Core State Standards and the *GAISE* recommendations.

Teachers' Understanding of Inferential Reasoning

Inferential reasoning is a complex process, which requires both conceptual knowledge of statistics and reasoning abilities. However, empirical studies investigating the knowledge needed for teaching statistics to middle and secondary school students are almost nonexistent (Groth, 2007; Shaughnessy, 2007). Leavy (2010) conducted a study specific to elementary preservice teachers, and found that most preservice teachers view mathematics and statistics *deterministically*, which means that tasks ostensibly have specific answers that are correct. However, in statistics, solutions account for variability contained in data sets. Therefore, all solutions also contain a degree of uncertainty. Given the disposition of elementary preservice teachers towards deterministic views of statistics, many are uncomfortable conducting classroom activities such as rolling dice in order to create an empirical data distribution that will not conform to theoretical expectations.

Research studies focused on preservice middle and secondary mathematics

teachers' knowledge of statistical inference are also limited. In a study of two preservice high school teachers with extensive mathematical backgrounds, Liu and Thompson (2005) found that transferring formal statistical knowledge regarding sampling distribution and confidence intervals to analyzing public opinion polls is nontrivial and problematic. Similarly, Heid, Perkinson, Peters and Fratto (2005) found that preservice secondary mathematics teachers have difficulty applying sampling distributions to real data and tend to reason deterministically rather than probabilistically when comparing sampling distributions to populations. In both studies, researchers assert that teacher preparation programs are not adequately preparing future teachers in terms of being able to generate inferences with real-world contexts and data, and more specifically, to understand the complex relationships of sampling distribution in comparison to populations.

Although several studies have explored preservice teachers' statistical knowledge, few investigations have examined inservice teachers' statistical thinking. In regard to inservice middle and secondary mathematics teachers, Nicholson and Darton (2003) reported that high school mathematics teachers with limited knowledge of statistical inference are especially uncomfortable in reasoning about statistical concepts and explaining the relationship between random process and inference. Given the complexity involved with learning about statistical ideas and the process of inference, these findings are not surprising.

Makar and Confrey (2004) led a professional development effort designed to build inservice secondary mathematics teacher knowledge specific to formal inference through analyzing students' performance on state mandated assessments. While the

teachers were generally comfortable comparing distributions informally and with descriptive statistical measures such as mean, they struggled to reason about multiple types of variation both *within* and *between* data distributions. Furthermore, teachers were unable to effectively draw inferences when provided with sampling distributions via computer simulations. Makar and Confrey highlight that overuse of computer simulation activities might, in fact, promote the development of some misconceptions specific to sampling distribution, such as the notion that a sampling distribution is representative of the larger population.

In a study that found teachers' reasoning to be more advanced than students', McClain (2002) designed a professional development course for middle school mathematics teachers based on grade 7 student learning trajectories. McClain found that teachers offered more sophisticated responses than typical grade 7 students by employing multiplicative and proportional reasoning strategies. Grade 7 students often reason additively and compare data in terms of frequency. For example, a student may add two samples together to arrive at an estimation of the population and generalize characteristics of the population based on the total number of data points at each value. Multiplicative reasoning involves visualizing how samples might compare if the process of sampling were repeated many times in order to generalize to population characteristics. Proportional reasoning involves thinking about the values of a data distribution in comparison to each other rather than the number of data points on any particular value. Hence, although commonalities exist between student and teacher learning, learning trajectories for teachers will be different than student learning trajectories.

In a related study, Watson (2008) reported on a classroom design experiment conducted with a grade 7 teacher over the duration of three lessons. Watson designed classroom activities with the teacher to introduce the concepts of hypothesis testing, shape, skewness, center, spread, outlier, sampling and justification through the use of data sets in TinkerPlots. Watson found that the teacher was comfortable leading all activities with the exception of those related to inference and justification.

In summary, middle and secondary school teachers are least prepared to teach statistics and probability (CBMS, 2001). From the few empirically-based studies involving K-12 teachers, middle and secondary preservice and inservice mathematics teachers demonstrate knowledge of proportional reasoning, sampling, and core statistical concepts. However, challenges arise related to sampling distribution and comparison of sample mean to populations, with and without the assistance of computer simulation. In addition, applying formal knowledge and procedures to real-world scenarios is difficult for teachers and underscores the problems teachers experience in coordinating conceptions and generating data-based arguments. More broadly, both inservice and preservice teachers struggle to effectively generate inferences and tend to reason deterministically.

Characteristics of Tasks Used to Assess Inferential Reasoning

A synthesis of research studies focused on either assessing or developing students' inferential reasoning abilities informed the selection of inferential reasoning tasks to be included in this study (e.g. Bakker, 2004; Cobb, 1999; Garfield et al., 2007; Watson, 2002). Because preservice teacher participants did not have experience with formal approaches to inference prior to beginning of the statistics course, tasks were selected that could be solved with both informal and formal approaches. Statistics education researchers and educational psychologists advocate the need for tasks to be: ill-structured, open-ended, represented visually, and embedded within a relevant context.

Ill-Structured

Reasoning effectively to generate inferences requires prior knowledge of core statistical ideas and an understanding of the relationships between them (Garfield et al., 2007). Informal approaches to reasoning are needed when problems either do not align with known solution methods or are presented before students possess the knowledge of such methods. As Means and Voss (1996) state, "Informal reasoning assumes importance when information is less accessible, or when the problems are more open-ended, debatable, complex, or ill-structured, and especially when the issue requires that the individual build an argument to support a claim" (p. 140). Therefore, the tasks found in these studies tend to ill-structured and open-ended in nature.

Tasks that are ill-structured share four common traits according to Goel (2009): (a) require more than a single cycle to generate a response, (b) place few if any logical constraints on the solution, (c) are open-ended in nature with multiple potential solutions,

and (d) encourage an incremental solution that is refined in cycles with a low level of commitment during the beginning cycles.

Watson and Moritz (1999) asked students across grade 3 to 9 to compare the performance of two classes on a mathematics quiz given the scores of each in a line plot format. The class sizes were unequal, making proportional reasoning a requirement. In addition, the task meets the specifications provided by Goel (2009) by placing limited constraints on the problem solver, encouraging multiple approaches and solutions, and requiring more than a one-step process. The question posed by Watson and Moritz requires a data-based justification embedded in a context. Therefore, the problem is not only ill-structured in nature, but also demands that the students provided an inferential argument based on the data provided.

Watson and Moritz report that a group of students, who reasoned in multiple steps, tended to choose one of two solutions paths. One solution path consisted of visual comparisons of the distributions, while the other relied upon numerical approaches, such as calculating the mean. When student approach ill-structured problems, they generally progress through four phases: "problem structuring, preliminary design, refinement, and detailing. These phases differ with respect to the type of information dealt with, the degree of commitment to generated ideas, the level of detail attended to, and a large number of vertical transformations" (Goel, 2009, p. 3). Vertical transformation represents the deepening of an idea related to the problem or task posed. As ideas are fleshed out in more detail, students also become more committed to their solution strategy. The omission of one correct answer or lack of problem constraints is the key factor for encouraging inferential reasoning. Watson and Moritz describe an iterative

process that some students embarked upon to first compare measures of center, then to consider other characteristics of the data distribution such as skew or range, and finally to coordinate all possible data comparisons together to produce a detailed response. These steps and the associated deepening of detail regarding the final solution provide a view into students' reasoning beyond tasks that are highly structured, which seek a predetermined solution. The use of ill-structured tasks in statistics education research studies is fairly common (Bakker, 2004; Brase et al., 1998; Pratt et al., 2008; Saldanha & Thompson, 2002; Stohl & Tarr, 2002; Watson & Moritz, 1999). Since a goal of this research study is to assess the preservice teachers' reasoning, ill-structured tasks were selected for the assessments.

Open-Ended

The appearance of open-ended tasks in statistics education research is driven by the desire to require informal approaches to tasks (Bakker, 2004; Cobb, 1999; Cobb et al, 2003b, Garfield et al., 2007; Watson, 2002; Watson & Moritz, 1999). According to Leathman, Lawrence, and Mewborn (2005) open-ended problems "elicit reasoning, problem solving, and communication" (p. 413). Characteristics of high quality open-ended tasks include the involvement of significant mathematics, the potential to solicit basic to sophisticated and abstract responses, and a balance between too much and too little information. Clearly, the bounds of ill-structured tasks and open-ended tasks overlap to some degree as the descriptions of both include common characteristics.

Many teacher-researchers initially introduce open-ended tasks to hone students' thinking and reasoning about a situation. Through whole class discussion, the open-ended tasks become closed as taken-as-shared meanings develop. In one study, students

were asked to determine which of two ambulance service providers was better and provide justification for their reasoning (Cobb, McClain & Gravemeier, 2003). Through a whole class discussion, students determined that viewing the data in two equal groupings provided the needed information to make a decision. Hence, the initially open-ended task became closed through the instructional process of establishing norms for acceptable justification.

In another study, students were asked to list their daily activities and rank them in terms of most to least variable and again provide justification (Garfield et al., 2007). In this case, the teacher-researchers used the students' initial responses to again build classroom-level consensus around the concept of variability. Given that the assessments were completed individually, a taken-as-shared meaning was not developed at the whole-class level or at the teacher-student level initially. Therefore, the students were required to decide in many instance what the relevant aspects of each task were and what constituted acceptable justification. This decision making process parallels the requirements placed upon the preservice teachers as they completed tasks used in this study.

Visual Representations

The use of visual representations of data related to inferential reasoning tasks is common. Tasks involving small sets of data ($n < 30$) generally provide dot plots and bar graphs to depict the distribution (Garfield et al., 2007; Watson 2002; 2008; Watson & Moritz, 1999). In contrast, tasks utilized by Bakker (2004) require students to sketch dot plots on paper. Other researchers provide more elaborate visual representations and display options to students during studies with either larger sample sizes or that requested

students to determine characteristics of an underlying population distribution from sampling (Cobb, 1999; Cobb et al., 2003; Pratt et al., 2008; Stohl & Tarr, 2002). A few researchers asked students to create mental representations rather than use technology tools or sketches (Brase et al., 1998; Saldanha & Thompson, 2002). In all of these cases, by providing visual representations, the students' were encouraged to attend to global characteristics and relationships rather than focus on any one statistic.

Context

According to the authors of the *GAISE* (2007) recommendations, "In mathematics, context obscures structure. In data analysis, context provides meaning" (p. 7). Hence the use of context is the norm in statistics education and instructors commonly introduce data sets in relation to some real-world phenomena or situation. However, the way researchers use context in their tasks varies substantially. On one hand, several researchers have created problem scenarios familiar to students in an effort to increase accessibility and leverage prior knowledge and experiences (Bakker, 2004; Garfield et al., 2007; Watson & Moritz, 1999; Watson, 2002; Watson, 2008). For example, Bakker developed a context of children's weight with children of similar age to those involved in reasoning about the tasks. Bakker noted that many researchers avoid this potentially sensitive topic, but reported that students were able to generate distributions that aligned closely with actual data. Similarly, Watson created a sequence of tasks that were based on measures of actual students' heart rates and arm-span lengths. Finally, Garfield et al. asked students to create data sets based on their own daily activities. By creating data sets that are close to the knowledge and experiences of the students, the focus of the tasks is on the reasoning process and possibly avoids confusion from lack of prior experiences

and knowledge. On the other hand, some researchers formulate tasks based on real-world contexts. Cobb (1999) and Cobb et al. (2003) created a variety of real-world contexts such as ambulance response times, success of speed traps, effectiveness of AIDS treatments, battery life spans, SAT scores, response time versus alcohol intake, and carbon dioxide production over time. Cobb et al. explain that prior research findings required that students find the context both plausible and of importance before they will engage in reasoning about the data. Therefore, the context of a given task should be discussed thoroughly prior to students embarking on any exploratory data analysis or inferential reasoning.

Finally, several researchers couch tasks in more traditional contexts based on probabilistic situations (Pratt et al., 2008; Stohl & Tarr, 2002). Stohl and Tarr (2002) posed problems in contexts involving dice, coins, marbles, and fish populations. The tasks offer a game-like feature and pose challenges to the inquisitive student. However, the contexts are not necessarily familiar to the students or realistic in nature.

The contexts for generating inferences abound that are both familiar to students and potentially relevant such as differences in bedtimes, allotted television viewing time, and distances from homes to school. Finding contexts that are familiar to students have the potential to stimulate rich and productive classroom discussions. Problems that are relevant and of interest to students may promote authentic engagement in data analysis activities through drawing inferences (Cobb et al., 2003). Garfield, delMas and Chance (2007) report that while they conscientiously selected contexts that would be familiar to students when designing tasks for a particular lesson, students were not interested in addressing the question posed by the tasks. Therefore, the context and question being

addressed must be both relevant and of interest to the student. In selecting tasks with these attributes, students will begin activities with some level of experiential knowledge and interest, which will increase the likelihood of building structural conceptions and mental models needed for inferential reasoning (Means & Voss, 1996).

The researchers who created the informal inferential reasoning tasks used within this study have charted new territory. The review of research strongly suggests that investigations into teachers' inferential reasoning should utilize tasks with several key characteristics: ill-structured, open-ended, represented visually, and embedded within a relevant context.

Summary

Inferential reasoning has recently become a prominent component of the grade 6-12 mathematics curriculum as determined by the Common Core State Standards (NGA Center & CCSSO, 2010). Research related to the process of inferential reasoning highlights the complexity in generating an inference due to the need for prior knowledge of statistical conceptions, an understanding of how the conceptions are related, and an ability to develop an argument that is supported by data and reasonable for the task context. In addition, misconceptions related to probabilistic reasoning pose additional challenges that must be overcome by many students and teachers alike. Therefore, statistics educators and researchers recommend that students begin learning how to informally generate inferences in middle grades and progress to formal approaches in secondary schooling.

Research related to the knowledge needed to teach inferential reasoning is indeed scarce (Shaughnessy, 2007). From the few studies that have been conducted, teachers

tend to be confident in their knowledge and teaching of many core statistical concepts. However, the process of generating inferences poses challenges for middle and secondary mathematics teachers, and they are uncomfortable teaching statistical content beyond core conceptions. This finding is unsurprising given the lack of teacher preparation related to probability and statistics.

The present study examines how a cohort of preservice middle and secondary mathematics teachers' inferential reasoning changes as they progress through a statistics content course designed specifically for future teachers. Recently, researchers have embarked upon studies related to how students' can be taught to reasoning inferentially and also how to characterize their inferential reasoning. In these studies, students' inferential reasoning is both assessed and fostered through the use of tasks that are designed to be: ill-structured, open-ended, visually represented and embedded in a context. Therefore, the tasks within this study embody these attributes.

In summation, teaching inferential reasoning to both preservice teachers and grade 6-12 students is a daunting challenge given the lack of curricular materials, unsuccessfulness of prior initiatives at the college level, deterministic dispositions of teachers and students, and inherent misconceptions related to probability. However, by reviewing the extant research related to inference, insights for understanding the underpinnings of inferential reasoning have been identified, including tasks for assessing inferential reasoning.

In the next chapter the research design and methodology for addressing the three research questions posed are discussed. First, I provide an overview of the context of the study and five data sources. Next, data collection processes are detailed for each source

including the timing and frequency of data collected for each source. I then describe how the data are analyzed, including how codes were assigned and aggregated to identify changes in the preservice teachers inferential reasoning and the opportunity to learn inferential reasoning. Finally, I offer limitations of the research design and methodology followed by a summary of the chapter.

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

The review of theoretical and empirical literature suggests additional studies are needed in order to better understand changes in inferential reasoning and preservice teachers' opportunities to learn key statistics content. Accordingly, the goal of this study is to characterize how middle and secondary preservice teachers' inferential reasoning changes during a statistics content course and analyze the associated opportunity to learn inferential reasoning the course affords. To achieve the desired objectives, the study employs both quantitative and qualitative research methods and leverages existing cognitive frameworks for data analysis.

In this chapter, I address the research methodology in three sections. In the first section, I present the research design of the study and include a description of the research setting, the participants, data sources and data collection processes. In the second section, I focus on data analysis, which contains information related to the: (a) cognitive framework used to characterize the cohort's inferential reasoning, (b) framework used to determine the cohort's opportunity to learn, (c) procedures used to tier task-level characterizations into broader categories, (d) reliability of the coding schemes, and (e) limitations of the study. In the final section, I provide a summary of the research methodology.

Research Methodology

Research Context

The setting for this research is the campus of a large, public, research university located in the Midwestern portion of the United States. Enrollment at the university

consists of approximately 31,000 students with 24,000 in undergraduate studies and 7,000 in graduate programs. Enrollment in the College of Education is approximately 3,000 students with nearly 800 certified teachers graduating in a given year, approximately two to three dozen of which are licensed to teach middle and/or secondary mathematics.

In order to fulfill requirements of the teacher development programs, preservice middle and secondary mathematics teachers must pass either the statistics course that is the focus of this study or a calculus-based version of the statistics course; however, few students opt for the calculus-based option. Typically, middle school mathematics preservice teachers are also required to take a prerequisite elementary statistics course that introduces core statistical concepts and teaches basic statistical reasoning processes; this beginning statistics course is also required of preservice elementary teachers. Exactly 75% of the preservice middle school mathematics teachers completed the prerequisite elementary statistics course prior to participating in this study.

The content of the statistics course under study reflects a structure typical of many introductory statistics courses, and includes topics such as descriptive statistics, permutations and combinations, probability, probability distributions, sampling, estimation, confidence intervals and hypothesis testing. The statistics class met one day per week for three hours in the evening. The format of the class sessions tended to be similar in nature each week with a short period of time allotted to answering students' questions regarding homework assignments followed by a more lengthy teacher-directed lecture of new content. The final third of the class time was generally dedicated to a small group activity or reserved for in-class assessments. The primary textbook for the

course was *Introduction to Probability and Statistics* (Mendenhall, Beaver & Beaver, 2005). The textbook contained a CD-ROM with data sets, applets and simulations to supplement the paper and pencil activities. MINITAB software was used to complete several small projects that were assigned as part of weekly homework sets. The in-class group activities were often assigned from the *Quantitative Literacy Series* (Mrdulla et al., 1995) with an emphasis placed on how the preservice teachers would modify the tasks for use in their own classrooms. These group projects were normally completed in class followed by a short discussion of findings. Course grades were based on homework, attendance, group projects, assessments and a culminating paper.

It is worth noting that the instructor of the statistics course held a Master's degree in statistics, a doctoral degree in an education field, and had experience in teaching secondary school mathematics, including teaching statistics. The instructor had taught this particular course several times, and therefore anticipated which aspects of the course content would pose challenges for the preservice teachers. Initially, 34 students were enrolled in the course, but 1 preservice teacher withdrew during the first few weeks of the semester. The course was held in a small auditorium with three writing boards. Preservice teachers sat in rows of tables, in chairs that were mounted to the tables. The secondary school preservice teachers tended to sit on one side of the auditorium with the middle school preservice teachers on the other and seldom interacted or worked together.

Participants

Subject selection. The subjects of this research study were selected based on their status as preservice middle and secondary school teachers enrolled during the spring semester of 2010 in the statistics course required for both populations. Of the 33

preservice teachers enrolled in the course, all agreed to participate in the study. Sixteen intended to teach middle school mathematics and 17 sought teaching licensure in secondary mathematics. The selection of this cohort of preservice teachers potentially affords unique insight into mathematics teacher learning, as few colleges or universities currently offer content-specific courses designed specifically for the needs of future teachers (Shaughnessy, 2007). A summary of the cohort's demographic profile data is provided in Table 3.1.

Table 3.1

Participants' Academic Standing and Gender Profile

Participants	Total	Sophomore	Junior	Senior	Female	Male
Preservice Middle School Mathematics Teachers	16	1	8	7	15	1
Preservice Secondary Mathematics Teachers	17	1	8	8	13	4

The cohort of preservice teachers was almost evenly split between middle and secondary school certification areas. In addition, most preservice teachers were either juniors or seniors and had completed at least four collegiate mathematics courses prior to this study. The cohort consisted mainly of females (85%) with males comprising a minority (15%) of the sample. Ethnicity data were not collected from the cohort due to the noticeable lack of diversity within the cohort.

Background coursework. Table 3.2 summarizes the cohorts' statistics coursework completed prior to the study. The cohort of preservice teachers had differing amounts of statistics coursework prior to participating in this study. All but one of the

middle school preservice teachers had completed a college-level statistics course previously. In comparison, only six of the secondary preservice teachers had completed similar coursework.

Table 3.2

Prior Tertiary Statistics Coursework Completed

Participants	Total	AP Statistics	Prerequisite	Other Statistics Course
Preservice Middle School Mathematics Teachers	16	1	12	2
Preservice Secondary Mathematics Teachers	17	2	3	1

However, the mathematics coursework completed by the secondary preservice teachers was considerably more rigorous in nature. The secondary program requires preservice secondary mathematics teachers to successfully complete a three-course sequence in calculus, whereas the middle school program requires merely one course in calculus. With the exception of AP statistics, the cohort's experiences with statistics during middle and secondary school mathematics coursework (e.g. linear regression during Algebra 2) was not collected as part of this study.

Accessing participants. During the first week of class, *all* preservice teachers enrolled in the statistics course provided written consent to participate in the research study. In particular, participants were informed that the research study would have no direct bearing on grades earned in the course.

Data Sources

Five data sources were used to characterize how the preservice teachers' inferential reasoning changes during a statistics course, explore the relationship between the changes in informal and formal approaches to drawing inference, and describe the cohort's opportunity to learn inferential reasoning: (1) pre-assessments, (2) midcourse clinical interviews, (3) post-assessments, (4) artifacts: written tasks used in the course, and (5) field notes of tasks written on the board during lectures and audio recordings of questions posed by the instructor. Each data source is listed in Table 3.3, including a description of the instrument, when data were collected, the frequency of collection, and the associated purpose.

Table 3.3

Data Sources: Timeline, Frequency and Purpose

Data Sources	Data Collection Timeline	Number	Purpose of the Data Source
Pre-assessments	Week 1 of class	33	To create a baseline of inferential reasoning ability for the cohort prior to further statistics instruction.
Midcourse clinical interviews	Weeks 6-8 of class	12	To provide information regarding the change in inferential reasoning midway through the course.
Post-assessments	Week15 of class	33	To provide information regarding the change in the cohort's informal and formal inferential reasoing.
Artifacts	During each class session	315 tasks	To describe the statistics covered and the emphasis placed on reasoning during the course.
Field Notes and Audio Recordings	During each class session	60 tasks	To describe the statistics covered and the emphasis placed on reasoning during the course.

I now describe the process I used to create written assessments and then offer a brief discussion of each assessment. Next, I describe the collection of artifacts, classroom audio recordings and field notes.

Designing the assessments. The primary sources of data for characterizing the change in participants' inferential reasoning are three written assessments. The assessments were administered at the beginning, middle, and end of the semester, and each consisted of seven to eight parallel tasks[1]. The administration of parallel tasks multiple times throughout a course, is advocated by Zieffler et al. (2008), who conducted a teaching experiment designed to develop college students' inferential reasoning during an introductory statistics course:

> These tasks (or parallel versions of the tasks) could be given to students at multiple times throughout a course or unit of instruction to examine how students' reasoning develops. This would allow instructors to examine how students use their informal knowledge and informal reasoning to draw conclusions and make inferences as they experience instruction related to informal or formal methods of statistical inference. (p. 52)

In this study, all tasks were purposefully selected from published research studies and statistics education materials and contain a balance of the two main categories of inferential reasoning: (a) comparison and determination of cause from randomized comparative experiments, and (b) generalizing from samples to a population. The demands of the tasks aligned with the process and content goals outlined in the *GAISE*

[1] The tasks on each assessment are *parallel* in terms of several factors. The contexts of tasks are balanced between real-world scenarios and settings specific to college students. In addition, the ratio of types of inferential reasoning tasks, reasoning from a sample to an unknown population and comparison and determination of cause from randomized comparative experiment, are common across assessments. Finally, the types of data representation formats are common across each assessment.

report for middle and secondary students (Franklin et al., 1997) and embody characteristics thought to promote inferential reasoning.

In addition to inferential reasoning tasks, two prerequisite knowledge tasks are included in the pre- and midcourse assessments. As noted previously, 20 of the 33 preservice teachers had completed a college-level statistics course focused on statistical concepts and reasoning prior to this research study. Therefore, many of the preservice teachers potentially possessed a portion of the needed knowledge and skills to successfully engage in informal inferential reasoning prior to participation in this study. In order test this assumption of prerequisite knowledge and reasoning ability, two tasks specific to statistical concepts and reasoning were included in the pre- and midcourse assessments. The prerequisite tasks assessed knowledge of measures of center, measures of variance, skewness, range, outliers, data distributions, and comparison of two complete populations.

The selection of inferential reasoning tasks to be included in the three assessments was informed by a synthesis of research studies focused on either assessing or developing students' inferential reasoning abilities (e.g. Bakker, 2004; Cobb, 1999; Garfield et al., 2007; Watson, 2002). Researchers advocate the need for tasks to be ill-structured, open-ended, represented visually and embedded within a real-world context.

In order to encourage informal reasoning versus procedural fluency, tasks are purposefully ill-structured and open-ended. Given the sequencing of statistics content in the course, the procedures for generating formal inferences were not taught until the final two weeks of class. Therefore, the cohort needed to employ informal methods, with the possible exception of three students who had completed an advance placement course in

statistics, to generate inferences during the pre-assessment and midcourse clinical interviews. According to Means and Voss (1996), "Informal reasoning assumes importance when information is less accessible, or when the problems are more open-ended, debatable, complex, or ill structured, and especially when the issue requires that the individual build an argument to support a claim" (p. 140). For example, one ill-structured task required participants to choose the best ambulance service based on sample dot plots of ambulance response times for two different companies (Figure 3.1).

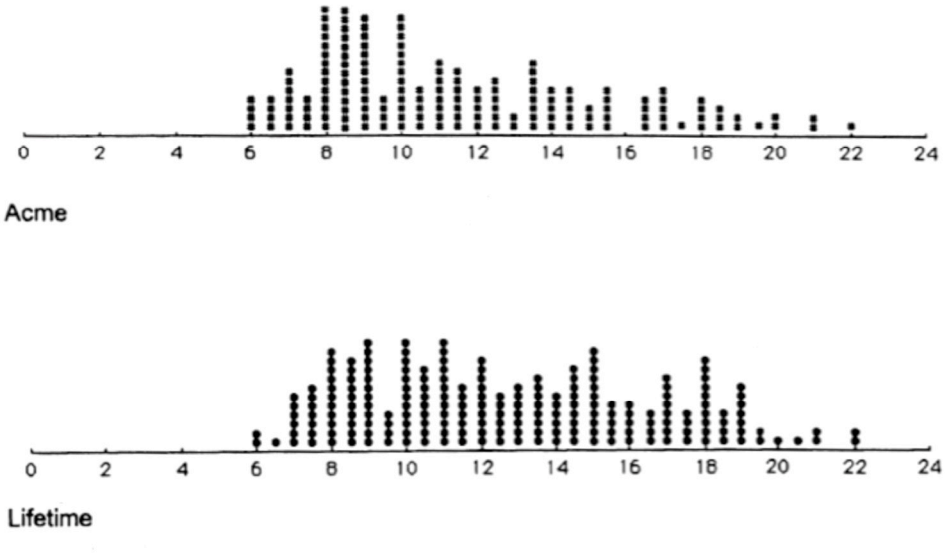


Figure 3.1. *Sample item, Ambulance Service task.* From "Learning about Statistical Covariation" by P. Cobb, K. McClain, and K. Gravemeijer, 2003, *Cognition and Learning, 21*(1), p. 26.

The decision of what it means to be the "best ambulance service" is left to the individual. Perhaps the provider who is most reliable will be selected or the one that is the fastest on average or a combination of factors will be considered. Therefore, the participant must interpret the problem statement of the task.

Once the problem has been defined more concretely in the mind of the participant, a determination of solution must also be made to bring closure to the open-ended nature of the task. In the case of the Ambulance Service selection task, the problem solver may

decide to base the decision solely on the average response time. At this point, a determination regarding the degree of difference must be made. For instance, if the average response times differed by only 1 minute, then can the services be judged comparable? Hence, the problem solver must reason yet again on the *practical significance* of findings and the appropriateness of the data to justify a particular stance.

To encourage reasoning that considers multiple factors, most of the assessment tasks used in this study employ visual representations of data versus tables or summary statistics. The purpose of visual representations is to draw the problem solver's attention to global characteristics and relationships rather than focus on any one statistic, such as the median or average (Bakker, 2004). For example, in the case of the Ambulance Service task, one provider may have a few outliers that are simply unacceptably high response times. Given the choice between slightly higher average response times or lower response times with greater variability, individuals may choose reliability over the possibility of receiving extremely slow ambulance response times. These high response times or outliers become invisible through the computation of average, but are readily apparent through a visual representation of response times. In addition to reasoning with given visual representations, several tasks ask for visual representations to be created from mental images or through the use of simulations.

Lastly, the use of a real-world context is common in all tasks and is generally the norm in statistics education. Approximately half of the assessment tasks are set in real-world scenarios similar to problems found within the preservice teachers' textbook or in a current events magazine. The remaining half of the tasks are set in contexts intended to be familiar to the preservice teachers and specific to college students. The familiar

contexts ideally increase accessibility and leverage prior knowledge and experiences (Bakker, 2004; Garfield et al., 2007; Watson & Moritz, 1999; Watson, 2002; Watson, 2008). By creating data sets that are close to the knowledge and experiences of the cohort, the focus of the tasks can be on reasoning processes and possibly avoid confusion from lack of prior experiences and knowledge. The context of these tasks may also be real-world in nature but are specific to the participants in this study. Table 3.4 provides a summary of the tasks used in the three written assessments.

Table 3.4

Task Summary for the Assessments

Task and Context	Statistical Concepts	Generalization from Samples	Comparative Experiments	Source
Pre-assessment				
1. Bedtimes	√			Author
2. Class scores	√			Watson, 1999
3. Weight		√		Bakker, 2004
4. Migraine treatments			√	Bright et al., 2003
5. Training programs			√	Zieffler et al., 2008
6. Diet and cholesterol			√	Cobb et al., 2003
7. Review session		√		Zieffler et al., 2008
Midcourse Assessment				
1. Class scores	√			Watson, 1999
2. Speed trap			√	Cobb, 1999
3. Ambulance service		√		Cobb, 1999
4. Cuckoo's Eggs		√		Reading & Reid, 2006
5. Diet and cholesterol			√	Cobb et al., 2003
6. Pennies and mints		√		Fong et al., 1986
7. Discrimination case		√		Burrill et al., 2003
Post-assessment				
1. Migraine treatments			√	Bright et al., 2003
2. Weight		√		Bakker, 2004
3. Training programs			√	Zieffler et al., 2008
4. Diet and cholesterol			√	Cobb et al., 2003
5. Review session		√		Zieffler et al., 2008
6. Speed trap			√	Cobb, 1999
7. Discrimination case		√		Burrill et al., 2003
8. Ambulance service		√		Cobb et al., 2003

Pre-assessment. Consisting of seven tasks, the pre-assessment serves as the initial data collection point regarding preservice teachers' inferential reasoning abilities prior to receiving formal instruction during the statistics content course. As previously mentioned, the first two tasks serve to validate prerequisite statistical knowledge and reasoning ability of the cohort. The first task was written by the author and trialed in a

pilot study with grade 5 students. The second task was administered to a larger number of students from grades 3 to 9 in a research study (Watson, 1999). Tasks 3 through 7 focus on inferential reasoning and either originate from research studies or reside in the book *Navigating through Data Analysis in Grades 9-12* by Burrill, Franklin, Godbold and Young (2003). See Appendix A for the entire pre-assessment document.

Midcourse clinical interviews. The midcourse clinical interviews serve to characterize how a subset of preservice teachers' reasoning changed midway through the statistics class. The midcourse clinical interviews are different than the pre-assessment in several ways. Rather than conducting the assessment in class, the interviews were conducted individually outside of class to a subset of the cohort. Twelve preservice teachers were selected to represent the cohort: a group of four participants who performed at the lowest, middle and highest tiers of inferential reasoning on the pre-assessment were selected to represent a wide range of statistical thinking and essentially comprise a stratified, random sample. In addition, within each stratum of inferential reasoning, the subset contained a balance of middle school and secondary mathematics education majors. The process for analyzing the preservice teachers' performance on assessments is discussed more fully in the data analysis section.

By conducting the interviews outside of class, an opportunity existed to ask the preservice teachers clarifying questions regarding their inferential reasoning. In order to accommodate this additional discussion component, midcourse clinical interviews lasted approximately 10-15 minutes longer than in-class assessments. After preservice teachers completed each task in writing, they were solicited to explain their reasoning and clarify written responses. All midcourse interviews were audio recorded and transcribed to

capture the preservice teachers' explanations. Lastly, each participant received a $15 gift card as modest compensation for the additional time spent outside of class.

The midcourse clinical interviews contain tasks parallel in nature to the pre-assessment and also originate either from research studies specific to students' inferential reasoning or reside in the books *Navigating through Data Analysis in Grades 9-12* by Burrill, Franklin, Godbold and Young (2003) and *Navigating through Data Analysis in Grades 6-8* by Bright, Brewer, McClain and Mooney (2003). The midcourse clinical interview tasks were administered *prior* to formal instruction on formal methods for generating inferences but *after* descriptive statistics, probability, probability distributions and sampling. See Appendix B for the entire midcourse clinical interview document.

Post-assessment. The post-assessment serves as the final data point used to characterize the preservice teachers' change in inferential reasoning. The post-assessment contains tasks parallel in nature to the pre-assessment and midcourse clinical interviews, and focuses solely on inferential reasoning. Because the preservice teachers had been taught formal methods associated with drawing inferences prior to completing this assessment, summary statistics were provided on 4 of the 8 tasks in the event formal methods were employed. However, visual representations and other aspects of the tasks remained in tact. Therefore, the preservice teachers' were given the opportunity to solve the tasks with multiple approaches similar to the previous assessments. The cohort was permitted to use a formula sheet and a calculator to complete this assessment similar to the process specified by the professor for course assessments. The preservice teachers' completed the collection of tasks within 50 minutes. See Appendix C for the full post-assessment document.

Artifacts, audio recordings and field notes. The tasks included in homework sets, class projects, lecture and in-class assessments serve as another data source. Throughout the semester, a total of 375 tasks were gathered through course resources, field notes and audio recordings. Collectively, these tasks represent the preservice teachers' opportunity to learn statistical reasoning. Consistent with the conceptual framework, opportunity to learn is examined through two lenses: the *statistical content* covered in the course and the emphasis placed on *reasoning*. The statistical content contained within the tasks characterizes the content covered in the course, and the questions posed by the tasks determines the emphasis on reasoning.

Data Collection

Once the preservice teacher cohort provided consent to participate in the study, a pre-assessment was administered during the first week of class. The purpose of the initial assessment was to formulate a baseline characterization of the cohort's inferential reasoning *prior* to instruction. During the middle of the course, clinical interviews were conducted with 12 preservice teachers, selected to represent a wide range in levels of performance on the pre-assessment as well as a balance of middle and secondary mathematics emphases. The purpose of the mid-course assessment was to capture changes in informal inferential reasoning prior to the introduction of formal inferential methods. In addition to completing tasks in writing, participants verbally explained their reasoning before proceeding to the next item. Finally, a post-assessment was administered to the entire cohort at the end of the semester. The purpose of the post-assessment was to establish a characterization of inferential reasoning *after* all the statistics content had been taught in the course.

I attended every class meeting to collect data regarding the statistical content covered and the emphasis placed on reasoning. In order to capture the questions posed by the professor, I audio recorded all lectures. Moreover, during each class session, I took field notes with particular attention on the tasks used during the instruction. Additionally, I collected all homework assignments, small group projects, and assessments for subsequent analysis.

Data Analysis

The first step in data analysis was to categorize the preservice teachers' responses to tasks included in the pre-assessment, midcourse clinical interview, and post-assessment. The cohort's informal responses to inferential tasks were characterized with the aid of a modified SOLO taxonomy based on the work Biggs and Collis (1982) and augmented by the research results of Mooney (2002). Formal responses were characterized through the application of the general SOLO taxonomy developed by Biggs and Collis. After completing the SOLO taxonomy characterization, each response was also coded for core and aggregate concepts used during reasoning. The second step was to aggregate the results into broader categories both at an individual and cohort level. In order to address the second research question, an investigation for relationships between the change in informal and formal reasoning was conducted using correlation techniques to measure the association. Finally, the preservice teachers' opportunity to learn was described through an analyses of the content covered in the course mapped against the components of the conceptual framework, and the extent to which reasoning was emphasized as determined by the intertwined mathematical strands of proficiency (Kilpatrick, Swafford & Findell, 2001).

Characterization of Informal Inferential Reasoning Responses

The augmented SOLO taxonomy described previously consists of a one-cycle SOLO taxonomy with informal inferential reasoning descriptors for each response type. Given Shaughnessy's (2007) statement that students and teachers often experience the same challenges when learning statistics, the descriptions of middle school students' informal inferential reasoning are applicable and align with the concrete symbolic mode of reasoning within the SOLO taxonomy. Details of the cognitive framework for characterizing informal inferential reasoning responses to tasks are provided in Table 3.5.

Table 3.5

Informal Inference Characterization Framework

Characterization of Responses	
Prestructural (P)	General description: The task is engaged, but the learner is distracted or misled by an irrelevant aspect. Inferential reasoning description: Makes inferences that are not based on the data or context.
Unistructural (U)	General description: The learner focuses on the relevant domain, and picks up one aspect to work with. Inferential reasoning description: Makes inferences that are primarily based on the data through a single correct comparison or a set of partially correct comparisons within or between data displays or sets. Some inferences may be only partially reasonable.
Multistructural (M)	General description: The learner picks up more and more relevant or correct features, but does not integrate them. Inferential reasoning description: Makes partially reasonable inferences that are primarily based on the data and context through multiple correct comparisons within or between data displays and sets.

Relational (R)	General description: The learner now integrates the parts with each other, so that the whole has a coherent structure and meaning. Inferential reasoning description: Makes reasonable inferences based on data and the context through multiple correct comparisons within and between data displays and sets.
Extended Abstract (EA)	General description: The learner now generalizes the structure to take in new and more abstract features, representing a higher mode of operation.

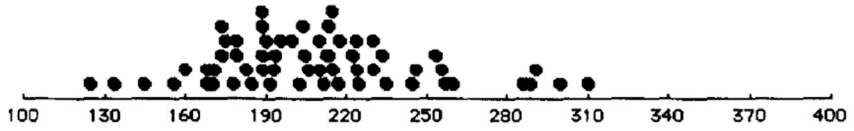
Note: Adapted from "A Framework for Characterizing Middle School Students' Statistical Thinking," by E. S. Mooney, 2002, *Mathematical Thinking and Learning, 4*(1), p. 37 and "Toward a Model of School-based Curriculum Development and Assessment Using the SOLO Taxonomy" by J. B. Biggs and K. F. Collis, 1989, *Australian Journal of Education, 33*(2), p. 152.

Application of the informal inference framework. Since formal approaches were not taught in the statistics course until after the *midcourse interviews*, the majority of preservice teacher responses to inferential reasoning tasks on assessments were informal in nature. Before a response was characterized, a determination was first made on the appropriate mode of learning according to the SOLO taxonomy (Biggs & Collis, 1982). For purposes of this study, response types consisted of informal approaches aligning with the concrete symbolic mode of learning or formal methods associated with the formal mode of learning. Occasionally, a preservice teacher mixed both approaches in a response to a task. In those cases, a dominant mode of response was sought prior to coding.

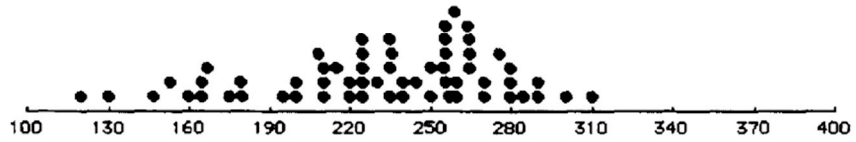
The four level descriptors for inferential reasoning within the cognitive framework relate to the preservice teachers' ability to generate and defend statistical inferences. The levels represent a cognitive progression through the concrete symbolic mode in a hierarchical fashion (Mooney, 2002). The first level, *Prestructural*, assumes that the inference generated in not based on the data. The remaining three levels assume that the preservice teachers offer a reasonable conjecture justified through data-based

argumentation appropriate for the context of the task. At the *Unistructural* level, reasoning will consist of a single idea or aspect of the data. At the *Multistructural* level, reasoning will consist of multiple ideas or aspects, but not in an integrated fashion. Finally, at the *Relational* level, the argument provided will relate conceptions in an integrated whole to provide a well-reasoned inference. The *Extended Abstract* level of response is equivalent to the *Prestructural* response level in the next mode. For this study, no responses were coded as *Extended Abstract.* For illustrative purposes, the Diet and Cholesterol task, which appeared on both the pre-assessment and post-assessment and hypothetical responses at each level of reasoning are provided in Figure 3.2.

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



After Program



Before Program

Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

Sample preservice teacher responses provided at each level of informal reasoning to the cholesterol and diet task contained within this study.

(P) *Prestructural*: No, lowering your cholesterol is not a happening of chance. This

occurred by eating more healthier foods that are less greasy.

(U) *Unistructural*: It definitely shows a shift in the clumping of the dots. Before the majority was between 220 & 280, but after it was clustered around 180 to 220.
[Core statistical concepts and aggregate statistical concepts: *center*]

(M) *Multistructural:* The two graphs show about the same range, so it possible that the program doesn't lower cholesterol. But for the most part I would say the program is effective. Before the program, most participants had a level of approx. 260. After the program, that number dropped to about 200. I think the program helps some, but not all participants.
[Core statistical concepts and aggregate statistical concepts: *spread, center*]

(R) *Relational* The majority of participants decreased their cholesterol levels. I suspect those participants who did not change much may have been affected by an outside factors, such as stress, failure to follow [the] program, etc. The mean dropped significantly, as did the median which went from 240ish to about 200.
[Core statistical concepts and aggregate statistical concepts: *variability, distribution, center*].

Figure 3.2. *Sample informal responses to the diet and cholesterol task.* From "Learning about Statistical Covariation" by P. Cobb, K. McClain, and K. Gravemeijer, 2003, *Cognition and Learning, 21*(1), p. 21.

Characterization of Formal Inferential Reasoning Responses

As mentioned previously, research specific to characterizing formal approaches to inference have not utilized the SOLO taxonomy. However, the SOLO taxonomy does accommodate a formal mode of learning, and researchers have identified the need to study how the SOLO taxonomy applies to more sophisticated learners and formal approaches in statistics (Watson et al., 1995). The initial coding of formal responses to inferential tasks followed the general SOLO taxonomy descriptions with the anticipated inferential reasoning descriptions provided in Table 3.6.

Table 3.6

Formal Inference Characterization Framework

Characterization of Responses	
Prestructural (P)	General description: The task is engaged, but the learner is distracted or misled by an irrelevant aspect. Inferential reasoning description: Makes inferences that are not based on the data or context.
Unistructural (U)	General description: The learner focuses on the relevant domain, and picks up one aspect to work with. Inferential reasoning description: Determines a formal approach for generating an inference, but applies the approach ineffectively with multiple errors. Inferences may be only partially reasonable or complete.
Multistructural (M)	General description: The learner picks up more and more relevant or correct features, but does not integrate them. Inferential reasoning description: Determines a formal approach for generating an inference, and applies the approach generally effectively with only minor errors. Inferences may be only partially reasonable or complete.
Relational (R)	General description: The learner now integrates the parts with each other, so that the whole has a coherent structure and meaning. Inferential reasoning description: Determines a formal approach for generating and inference, and applies the approach effectively without errors. Makes reasonable inferences based on data and the context.
Extended Abstract (EA)	General description: The learner now generalizes the structure to take in new and more abstract features, representing a higher mode of operation.

Application of the formal inference framework. As data were analyzed, the framework descriptions were refined to reflect the preservice teachers' actual reasoning progression. Recall that formal approaches necessitate the use of formulas, calculations

and tables in order to generate inferences. Therefore, responses that rely upon hypothesis tests, confidence interval computations, and similar methods were identified as *formal*.

Once again, the four level descriptors for inferential reasoning within the cognitive framework relate to the preservice teachers' ability to generate and defend statistical inferences. The levels represent cognitive progression through the formal mode in a hierarchical fashion (Biggs & Collis, 1989). The first level, *Prestructural*, assumes that the inference generated in not based on the data. Responses that identify the need for a statistical test, but progress no further were determined to be in the formal mode and *Prestructural*. Similarly, if a response identifies an inappropriate statistical test, the response was categorized as *Prestructural.* The remaining three levels assume that the preservice teachers choose appropriate types of formal approaches and generate inferences. At the *Unistructural* level, the application of the approach contains multiple errors and prohibits generating a reasonable inference. The focus of errors is not related to computational mistakes, but rather on entering wrong values for key variables such as the average, standard deviation or population parameters. Other noteworthy errors relate to incorrect *p*-values or confidence intervals. In addition, errors in interpreting the results of the formal methods may be present. Therefore, the focus is on understanding how the formal approaches apply to a task and what they mean once executed. At the *Multistructural* level, an appropriate method is selected and applied with only minimal errors. Finally, at the *Relational* level, the argument provided consists of an appropriate method, applied correctly, and an inference interpreted in relation to the task context. In addition to these SOLO taxonomy categorizations, all responses were coded for core statistical concepts and aggregate statistical concepts employed during reasoning. An

example of each response type is provided for the Hiring Discrimination task (Figure 3.3).

Task 7: Hiring of Managers and Discrimination
In 1972, 48 bank supervisors were each randomly assigned a personnel file and asked to judge whether the person represented in the file should be recommended for promotion to a branch-manager job described as "routine" or whether the person's file should be held and other applicants interviewed.

The files were all identical except that half of the supervisors had files labeled "male" while the other half had files labeled "female". Of the 48 files reviewed, 35 were recommended for promotion. Twenty-one (21) of the 35 recommended files were labeled "male", and 14 were labeled "female."

If the selection of the 35 candidates were purely fair in terms of gender given equal qualifications for promotion, we would expect that half the candidates would be male (17.5).

Question: As a member of a jury, would you confidently support a verdict that the bank supervisors discriminated against female candidates? Support your response.

Sample responses at each level of reasoning to the hiring discrimination task

(P) *Prestructural*: Yes, because $\frac{14}{48} = 29\%$. That means only 29% were females. That is a lot lower than 50% to where they could have discriminated.

Comment: The response recognizes the need for proportional reasoning, but progresses no further.

(U) *Unistructural*: $z = \frac{12 - 17.5}{1.6/\sqrt{35}} = \frac{3.5}{1.6/\sqrt{35}} = 12.9$

I would support the verdict that the bank supervisors discriminated against female candidates.

Comment: The response selected a formal method of comparing the result to the expected outcome, but does not incorporate proportional measures. The inference generated is based on the data, but is not correct.

(M) *Multistructural:*

$$\frac{21}{35}male, \frac{14}{35}female$$
$$= 17.5, \sigma = 1.65$$
$$z = \frac{(\frac{21}{35} - \frac{14}{35})}{\sqrt{\frac{(.6)(.4)}{35} + \frac{(.4)(.6)}{35}}} = 1.7078$$

No, I would not feel confident in supporting a verdict that the bank discriminated against female candidates.

Comment: The response utilizes both a formal method and proportional measures to compare the outcome to the expected value with only minor errors. The inference generated is reasonable based on the data and context.

(R) *Relational:*

$$z = \frac{(\hat{p}_1 - p_0)}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{(\frac{21}{35} - \frac{17.5}{35})}{\sqrt{\frac{(.5)(.5)}{35}}} = 1.18$$

There is no statistically significant evidence even at the .10 level of significance that the true proportion of men being recommended for promotions is not actually 0.5 (equal chances of being promoted as a female). This sample may have occurred by chance.

Comment: The response contains the appropriate statistical model and associated values. The inference generated is based on the data and context and demonstrates an integrated understanding of the formal methods and procedures.

Figure 3.3. *Sample formal responses to the hiring discrimination task.* Adapted from *Navigating through data analysis in grades 9-12* by G. Burrill, C. A. Franklin, L. Godbold, and L. C. Young, (2003). National Council of Teachers of Mathematics, Reston, Virginia, p.104.

Dominant Level of Reasoning

The dominant levels of inferential reasoning for each preservice teacher on each assessment were determined by converting the level of responses to integer values. The values were 0, 1, 2, and 3 corresponding in order to the response types of P, U, M and R. For each preservice teacher, the arithmetic mean was computed and translated to a dominant level of reasoning on each assessment. When an average fell equally between

reasoning levels, the response ratings of P, U, M and R were viewed to determine the appropriate dominant level of reasoning by finding the modal level of reasoning between the two adjacent levels. For example, if a preservice teacher's level of reasoning fell between the levels of Unistructural and Multistructural reasoning, the individual task reasoning levels were viewed to identify the modal level of reasoning between these two adjacent levels; that is, analysis was undertaken to determine whether Unistructural or Multistructural reasoning was more common. If Unistructural responses occurred most often on the assessment, the dominant level of reasoning was classified as Unistructural.

It is worth noting that one task, appearing on both the pre- and post-assessments, was ultimately excluded from the analysis because of its potentially confounding results. Specifically, Task 7 on the pre-assessment (Appendix A) and Task 5 on the post-assessment (Appendix C) was embedded in a context of assessing the effectiveness of a review session. The majority of preservice teachers responded to this task at the Prestructural level on both the pre-assessment (73% of the cohort) and post-assessment (52% of the cohort). The context of the task appeared to override thinking to such an extent that most participants *ignored* the data provided. Even though the task specified that the students who attended the review session were *randomly selected*, many respondents made no reference to the data but instead imposed their own interpretations as evidenced by comments such as, "Students who attend the review session might have been more likely to study more on their own as well, and I think that is more of the cause than strictly the review session." Responses to the Review Session tasks on both the pre-assessment and post-assessment were not representative of the quality of responses provided by the cohort. Preservice teachers, who otherwise reasoned in a consistently

sophisticated and statistically sound manner, predominantly reasoned at the Prestructural level on this particular task. Since the responses to this task on both assessments were determined to not be representative of the cohort's inferential reasoning, this particular item was ultimately excluded from the data analysis.

Informal and Formal Inferential Reasoning

The relationship between the development of informal and formal inferential reasoning was examined through a rank correlational analysis. The preservice teachers' informal responses to inferential tasks were analyzed in comparison to formal responses both in absolute and relative terms. The characterization of response types of P, U, M, and R were converted to an interval scale as mentioned previously. This conversion reflects the hierarchical progression of learning within a mode (Biggs & Collis, 1989). A dominant mode of reasoning for the preservice teachers was determined for both informal and formal approaches to inference upon completion of the statistics course. These two measures were compared to see if a relationship exists between the two types of approaches to inferential tasks. Since formal methods were introduced during the later portion of the semester, the post-assessment task responses were the data source for this analysis. While it was expected that participants would provide formal responses to several tasks, four preservice teachers chose to reason exclusively with informal approaches. Hence, a formal inferential reasoning characterization was not created for these four participants given that their responses to assessment tasks were exclusively based on informal methods.

Opportunity to Learn

All tasks assigned for homework, projects or assessments were gathered and analyzed. In addition, tasks used during instruction were collected. These tasks provide a view of the preservice teachers' opportunity to learn throughout the course. In order to categorize the content, each task was assigned to one of the three content components shown in the conceptual framework: core statistical concepts, aggregated statistical concepts, and methods and procedures. Within each component, additional specificity was provided. Core statistical concepts include: measures of center, skewness, spread and variance. Aggregate statistical concepts include: distribution, sampling, and sampling variability. Finally, methods and procedures relate to all formal methods used to generate inferences and include: confidence intervals and hypothesis testing. Due to the high frequency of probability tasks, a category was created specifically for these tasks although not included in the conceptual framework. If a task covered a topic outside of these descriptions, a category called "Other" was designated with a description of the statistical content associated with completion of the task.

The questions posed in writing and verbally by the instructor regarding the tasks indicate the emphasis placed on reasoning throughout the course. The questions were categorized based on the mathematical proficiency strands described in the book *Adding it up* (Kilpatrick, Swafford & Findell, 2001). Each question was assigned to one or more of the following four categories: conceptual understanding, procedural fluency, strategic competence, or adaptive reasoning (Figure 3.4). Questions were not assigned codes specific to productive disposition.

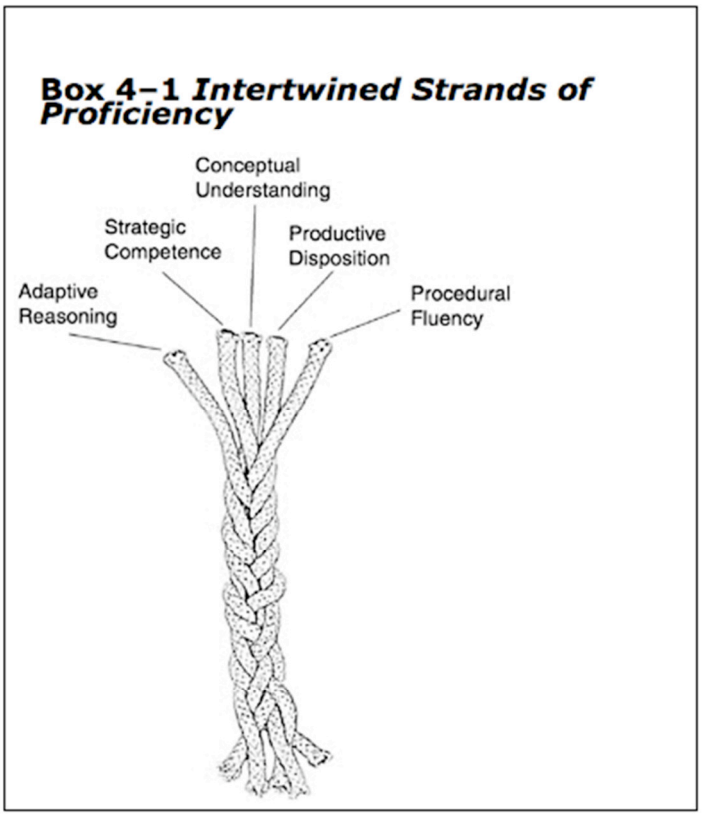


Figure 3.4. Strands of mathematical proficiency from *Adding It Up.* From *Adding it up: Helping children learn mathematics*, by J. Kilpatrick, J. Swafford and B. Findell, 2001, p. 117. Reprinted with permission.

Through this compilation of questions and tasks posed to the preservice teachers throughout the statistics course, the demands associated with statistical reasoning are viewed in comparison to other types of thinking requirements during the course. Tasks were coded for all content and mathematical strands included. Therefore, tasks generally received more than one statistical content code and one mathematical proficiency code. Tasks for the textbook often required multiple steps to be completed and spanned several content categories. In addition, tasks often asked one conceptual understanding question followed by several procedural fluency questions. In summary, the opportunity to learn data analysis creates a learning environment context for interpreting the changes in the cohort's inferential reasoning from the beginning to end of the semester.

Reliability

The reliability of the coding schemes for characterizing student responses on assessments and the description of course tasks is critical to the integrity and credibility of the study. Therefore, the director of my dissertation research was asked to characterize 20 of the preservice teachers' task responses from the pre- and post-assessments according to the cognitive frameworks described previously. The preservice teachers' tasks chosen for establishing reliability represent a stratified random sample with 10 tasks originating from the pre-assessment and 10 from the post-assessment. Through this process, 85% of the assessments were coded at exactly the same level of inferential reasoning, 10% differed by one level of reasoning, and 5% differed by two levels of reasoning.

In addition, the director of my dissertation research characterized 20 randomly selected course tasks in relation to statistical content and the emphasis placed on reasoning based on the mathematical strands of proficiency. With respect to the statistical content coding reliability, 90% of the course task codes matched completely, 5% differed by one content area, and 5% differed by two content areas. For the mathematical strands of proficiency, 95% of the course tasks codes matched completely with only 5% differing by one code.

Differences in codes for both assessment responses and course tasks were adjudicated and negotiated codes were used in subsequent analyses. The reliability checks occurred after the assessment data were collected and coded by the author. Through this design, categorization of the cohort's responses and course tasks received independent verifications to ensure reliability.

Limitations of the Study

The cohort in the proposed study resides within one statistics class, which may limit the generalizability of findings. However, because steps were taken to characterize the opportunity to learn specific to inferential reasoning, results viewed within a learning context can be interpreted more broadly.

Another potential limitation is common to studies that rely upon constructed-response assessments. The reasoning for the cohort is determined by what the preservice teachers actually communicated on paper in the case of the pre-assessment and post-assessment. The midcourse assessment results are more robust given the opportunity to verbally explain answers after each task was completed.

Lastly, the tasks within the course were algorithmic in nature and did not resemble the tasks on the assessments. Therefore, participants may have viewed the tasks as unrelated to the course content or different enough that transferring course learning proved to be difficult.

Summary

The research methodology and data analysis processes described enable a credible and detailed characterization of how the cohort's inferential reasoning changed throughout the semester and the associated learning context afforded by the statistics course. An overview of the research setting was described and conveys an ideal situation for data collection in that all preservice teachers willingly participated in the study and completed all requested assessments. In addition, the prior education and program requirements of the cohort suggest that preservice teacher participants were an able group in terms of their mathematical knowledge. The tasks, coding processes and frameworks

have been utilized in multiple research studies of educational settings in statistics, which lends credibility to findings and builds upon the existing knowledge base. The hierarchical nature of the SOLO taxonomy and multimodal learning approach provide flexibility in analyzing the relationship between informal and formal changes in inferential reasoning, and also enable a conversion of characterization measures to an interval scale for synthesizing findings both at an individual and cohort level. Research findings are offered in detail in the next chapter.

CHAPTER 4: ANALYSIS OF THE DATA AND RESULTS

Given the rise in prevalence of statistics in the mathematics curriculum for grades 6 through 12, there exists a pressing need to prepare preservice middle and secondary teachers to teach statistical content and processes. This study examines the statistical reasoning of preservice teachers related to inference, characterizing both the change in preservice teachers' reasoning ability and their opportunity to learn key statistical content. First, I present how the cohort of middle and secondary mathematics preservice teachers' inferential reasoning changed over the duration of a statistics course. Secondly, I explore the association between changes in the preservice teachers' *informal* and *formal* inferential reasoning. Lastly, I characterize the preservice teachers' opportunity to learn inferential reasoning.

Characterizing the Change in the Preservice Teachers' Inferential Reasoning

Results specific to characterizing the preservice teachers' changes in inferential reasoning are presented by examining the variation in responses to tasks at the class level and instability of inferential reasoning at the individual level. Next, dominant levels of inferential reasoning at the individual level are discussed for the time points relating to the beginning and end of the statistics course. Finally, the changes in the preservice teachers' inferential reasoning from the beginning to the end of the statistics course are characterized and illustrated through representative cases.

During the first week of the statistics course, the entire cohort (n=33) completed a pre-assessment consisting of seven tasks. Two prerequisite knowledge tasks, Bedtimes and Class Scores, were included in the pre-assessment (See Appendix A). The Bedtimes

task assesses preservice teachers' understanding of core and aggregate statistical concepts. Responses to the Bedtimes task were coded as either *correct* or *incorrect*, but not used to determine changes in preservice teachers' inferential reasoning. The Class Scores task responses were coded in the same manner as the inferential task responses and received either a Prestructural (P), Unistructural (U), Multistructural (M), or Relational (R) designation according to the modified SOLO taxonomy descriptions in Tables 3.5 and 3.6. Variation in the responses given by the preservice teachers was observed both at the *class level* for specific tasks and at the *individual level* across tasks.

Variation in Inferential Reasoning Responses at the Class Level

In order to determine the level of inferential reasoning on pre- and post-assessments for a specific task, coded responses were converted from an ordinal (categorical) scale to an interval (numerical) scale. Specifically, each level of cognitive reasoning was converted to an integer value of 0, 1, 2, or 3 for Prestructural, Unistructural, Multistructural, and Relational, respectively. In doing so, the mean and standard deviation for each task at the cohort level could be computed and compared across assessments.

All assessment tasks were specifically designed to elicit inferential reasoning and require evidence to support inferences. While the data representations and contexts of the tasks differed, responses to tasks at the cohort level tended to exhibit commonalities in terms of eliciting a wide range of responses. The standard deviation for tasks at the cohort level ranged from 0.35 to 0.99 (approximately one third to one level of inferential reasoning) on the pre-assessment, and from 0.79 to 1.08 (approximately one level of inferential reasoning) on the post-assessment. Hence, more variation in levels of

inferential reasoning was evident in responses to specific tasks on the post-assessment in comparison to the pre-assessment. Tasks with standard deviation values near 1.0 tended to elicit responses across the full spectrum of levels of inferential reasoning. Those tasks with standard deviation values below 0.75 tended to elicit primarily one level of inferential reasoning response for the majority of preservice teachers.

Standard deviation values below 0.75 were observed for two tasks on the pre-assessment, the Migraine Treatment and Training Programs tasks. On the Migraine Treatment task, preservice teachers exhibited the least variation with a standard deviation of 0.35 and a mean of 1.06, which equates to a Unistructural mode of inferential reasoning. In fact, 29 of 33 preservice teachers provided a Unistructural response to this task on the pre-assessment. Most preservice teachers only commented on the shorter response time of one medication versus the other. The faster response provided by one of the medications was visually apparent, and few participants commented on other attributes of the two distributions. This same task appeared on the post-assessment, but elicited a wider range of responses with only 17 of the 33 preservice teachers again providing a Unistructural response, a marked decline. In addition to comparing average response times, a portion of the cohort attended to other characteristics of the data distributions, such as range and spread. These more detailed and comprehensive responses increased the variation in levels of reasoning and concurrently increased the standard deviation from 0.35 on the pre-assessment to 0.79 on the post-assessment.

The second task that elicited a limited range of responses on the pre-assessment was the Training Program task, with a standard deviation of 0.69. A Unistructural level of reasoning was again common for this task with 20 of the 33 preservice teachers

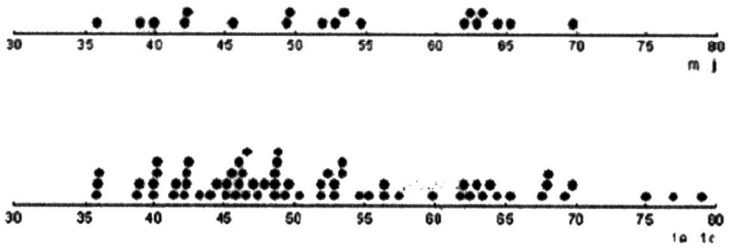
generating a response at this level of inferential reasoning. This same task appeared on the post-assessment and elicited a wider variation in responses with a standard deviation 0.85 and only 13 preservice teachers provided a Unistructural response.

The remainder of the tasks elicited responses at all four levels of inferential reasoning and did not produce one level of reasoning for the majority of preservice teachers. It is noteworthy that one task, the Review Session, was problematic and discarded from data analysis due to context. This task appeared on the pre- and post-assessments and was problematic in both cases. The context for this problem focused on the effectiveness of a review session in improving students' performance on an exam. This context evidently triggered biased and emotional responses, with few responding analytically. Preservice teachers' performance on this task was predominantly Prestructural on both assessments. The remainder of the tasks elicited a variety inferential reasoning responses and yielded similar levels of variation on both the pre- and post-assessments. The two tasks with lower levels of variation on the pre-assessment, Migraine Treatment and Training Programs, produced higher levels of variation on the post-assessment as the cohort attended to additional attributes of the data distributions and gained familiarity in interpreting box plot representations.

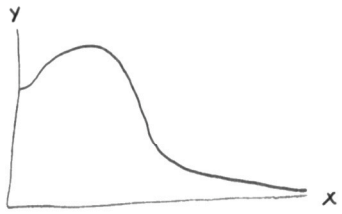
In order to illustrate the typical variation of inferential reasoning responses to a given task, archetypal responses of each level of coding are described for the Weight task on the pre-assessment, beginning with an example of a Unistructural (U) response in Figure 4.1

Task 3: Weight of Grade 7 Students

Below are two sets of real data, the first one with 27 values and the second one with 67, showing the weights (in kilograms) of grade 7 students from Columbia, Missouri.



Question: Based on this actual data, what would you estimate the shape of the distribution of 1000 grade 7 students' weights might look like? Please sketch below and label your axes.



Provide an explanation and rationale for your sketch above. There are many more students who are on the lower half of the range than there are on the higher half.

Figure 4.1. Example of a Unistructural response

The description of a Unistructural response is that "inferences are primarily based on the data through a single correct comparison or a set of partially correct comparisons within or between data displays or sets with some inferences possibly being only partially reasonable" (Table 3.5). In response above, the preservice teacher has identified that the larger population curve will be smooth in nature rather than comprised of individual data

points or bars of data. However, the respondent failed to recognize that the shape of the data distribution has two distinct mounds and also that students cannot have a weight of zero kilograms. Therefore, the response is categorized as Unistructural because the inference is based on a single correct comparison.

An archetypal Multistructural (M) response is shown in Figure 4.2.

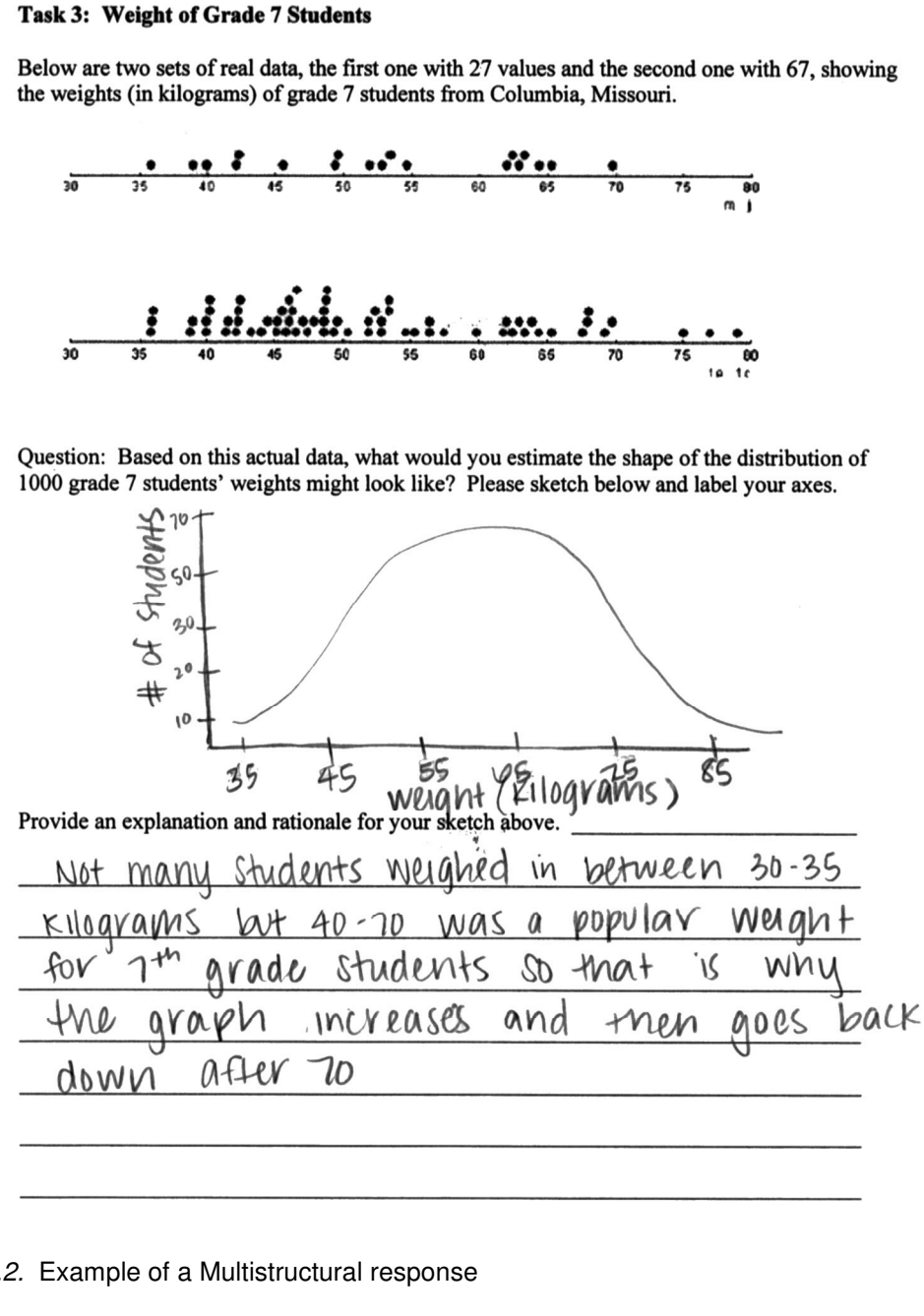


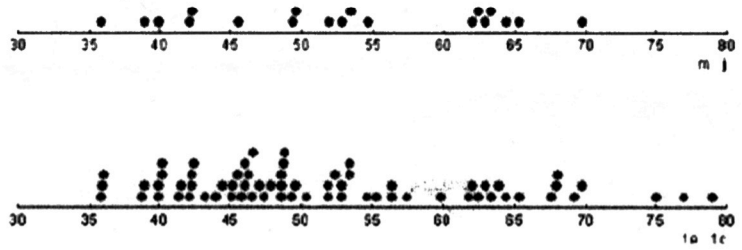
Figure 4.2. Example of a Multistructural response

The description of a Multistructural response is "a partially reasonable inference that is primarily based on the data and context through multiple correct comparisons within or between data displays and sets." In Figure 4.2, the preservice teacher has identified that the graph of the population will be smooth in nature and that the range of data will be similar to that of the two samples. The range of the population weights begins at 35 kilograms and trails off quickly after 85 kilograms. Therefore, the response reveals attention to multiple correct comparisons, but fails to recognize the shape of the distribution from the two samples. This response reflects a common misconception that all populations will resemble the normal distribution (delMas, Garfield, Ooms & Chance, 2007).

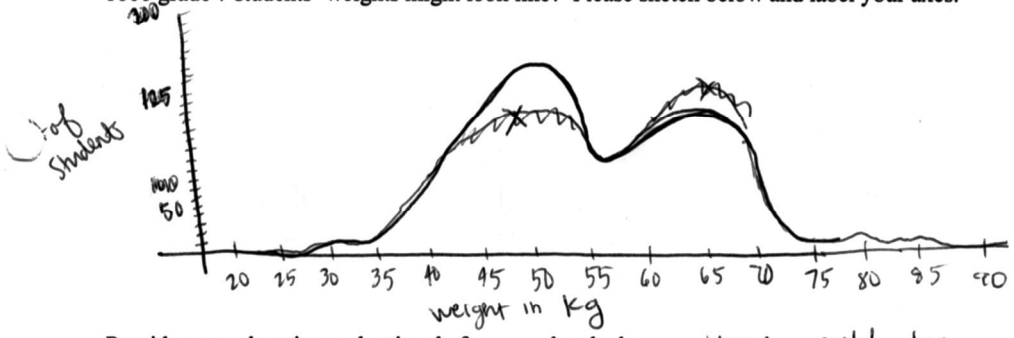
The most sophisticated level of reasoning is Relational (R), and an archetypal response of such is depicted in Figure 4.3.

Task 3: Weight of Grade 7 Students

Below are two sets of real data, the first one with 27 values and the second one with 67, showing the weights (in kilograms) of grade 7 students from Columbia, Missouri.



Question: Based on this actual data, what would you estimate the shape of the distribution of 1000 grade 7 students' weights might look like? Please sketch below and label your axes.



Provide an explanation and rationale for your sketch above. there would be several outliers in the high 20s & in the 70s or 80s. The greatest portion of population is between 35 and 55 but again a lot between 60 and 70.

Figure 4.3. Example of a Relational response

The description of a Relational response is "makes reasonable inferences based on data and the context through multiple correct comparisons within and between data displays and sets." In Figure 4.3, the preservice teacher has incorporated the (a) range of data, (b) shape of the data distribution, and (c) smoothness associated with a large

population. In addition to the graph, the written response acknowledges the characteristics of the samples and how this information is used in generating an inference related to the population. Relational responses provide a comprehensive inferential argument using all relevant aspects of the data provided, whereas Multistructural responses present two relevant aspects often in a disjointed manner or unrelated to the context of the problem.

The aforementioned examples of preservice inferential reasoning at different levels depict the variation of responses observed for tasks on the pre- and post-assessments. Next, the relative stability of reasoning at the individual preservice teacher level across tasks is presented.

Instability of Inferential Reasoning Across Tasks

Pre-assessment. Using the same process for transforming responses from an ordinal to an interval scale, the mean at the cohort level on the pre-assessment is 1.27, which equates to an inferential reasoning level between Unistructural and Multistructural, but closer to Unistructural. Therefore, as a class the cohort reasoned slightly above a Unistructural level on the pre-assessment. The average standard deviation on the pre-assessment for the cohort was 0.84, or slightly less than one level of inferential reasoning on the pre-assessment with a range of 0.45 to 1.34 for individual preservice teachers.

The 11 preservice teachers, or one-third of the cohort, with the highest variation in responses across tasks ranged in standard deviation values from 0.96 to 1.34. This portion of the cohort tended to exhibit inferential reasoning responses at three or more levels across tasks on the pre-assessment. In order to categorize individual preservice teachers' reasoning on each assessment, a dominant level of reasoning was determined by

first computing the average reasoning level in numerical terms. For example, one preservice teacher provided two Prestructural responses, two Unistructural responses and one Relational response for an overall average of 1.0 on the pre-assessment, equating to the Unistructural level, with a standard deviation of 1.22. The preservice teacher was generally reasoning at either the Unistructural and Prestructural level. The presence of the Relational response supports the notion of Unistructural as a dominant mode as does the average value of 1.0. If the numerical value fell between two possible reasoning levels, response levels for individual tasks were viewed to determine which of the two adjacent levels occurred most frequently. Through this process, a dominant level of inferential reasoning was obtained for each preservice teacher on the pre- and post-assessments. For 9 of the 11 preservice teachers with the highest variation in responses on the pre-assessment, the dominant level of inferential reasoning was Unistructural, with one preservice teacher classified as Prestructural and another as Multistructural.

Another group of 11 preservice teachers demonstrated the least variation in responses, with standard deviation values ranging from 0.45 to 0.55. This subset of 11 preservice teachers provided responses at only two levels of inferential reasoning. As an example, one preservice teacher provided four responses at the Unistructural level and one at the Multistructural level. Therefore, the preservice teacher exhibited more stability in the level of inferential reasoning across tasks than those with who responded at three levels. For 10 members in this group, the dominant level of inferential reasoning was Unistructural and 1 was classified as Prestructural.

A summary of the dominant modes of inferential reasoning on the pre-assessment is shown in Table 4.1. Percentages in the table have been rounded.

Table 4.1

Dominant Levels of Inferential Reasoning on the Pre-Assessment

Participants	Prestructural Number (%)	Unistructural Number (%)	Multistructural Number (%)	Relational Number (%)
All Preservice Teachers	2 (6%)	23 (70%)	8 (24%)	0 (0%)
Middle School Mathematics Teachers	2 (13%)	11 (69%)	3 (19%)	0 (0%)
Secondary Mathematics Preservice Teachers	0 (0%)	12 (71%)	5 (29%)	0 (0%)

Most preservice teachers' initial inferential reasoning level was Unistructural in nature at the beginning of the statistics course. Middle school mathematics preservice teachers performed at slightly lower levels in comparison to their secondary school peers. Two middle school preservice teachers reasoned at the Prestructural level, most (11 of 16) at the Unistructural level, and three at the Multistructural level. In comparison, approximately two thirds of the secondary teachers reasoned at the Unistructural level with the other third reasoning at the higher level of Multistructural.

Post-assessment. During the last week of the statistics course, the entire cohort of preservice teachers completed the post-assessment (See Appendix C). All of the tasks required inferential reasoning. The mean inferential reasoning level for the cohort on the post-assessment was 1.7, which equates to an inferential reasoning level between Unistructural and Multistructural, but closer to Multistructural. Therefore, an increase in inferential reasoning was observed, with the cohort reasoning slightly below the Multistructural level. The average standard deviation for the cohort was 0.75 with a

range of 0.38 to 1.13 for individual preservice teachers, slightly less than was observed on the pre-assessment.

The 11 preservice teachers with the highest variation across tasks ranged in standard deviation values from 0.90 to 1.13 (approximately one level of inferential reasoning). This portion of the cohort tended to exhibit inferential reasoning responses at three or more levels across tasks on the post-assessment. For example, one preservice teacher provided one response at the Prestructural level, three at Unistructural level, two at the Multistructural level, and one at Relational for an average numerical response value of 1.4 and a standard deviation of 0.98. Since the average numerical value fell between the equivalent Unistructural and Multistructural levels of reasoning, the frequency of tasks coded to each of these two categories was compared in order to determine the dominant level of reasoning. In this case, more Unistructural level codes were assigned overall, resulting in a dominant mode of Unistructural. In terms of the individual dominant levels of inferential reasoning, eight of the preservice teachers in this subset performed at the Multistructural level with three at the Unistructural level.

A subset of 11 preservice teachers with the least variation across tasks ranged in standard deviation values from 0.38 to 0.58 on the post-assessment. This subset provided responses at only two levels of inferential reasoning, and therefore exhibited significantly more stability in their level of inferential reasoning than those with who responded at three or all four levels. As an example, one preservice teacher provided four responses at the Multistructural level and three at the Relational level for an average inferential reasoning value of 2.4 with a standard deviation of 0.53. Because more responses were present at the Multistructural, the dominant mode of reasoning for this preservice teacher

on the post-assessment was also determined to be Multistructural. Members of this subgroup demonstrated dominant inferential reasoning levels across the three highest levels with six at the Unistructural level, three at the Multistructural level, and two at the Relational level.

It is worth noting that changes occurred in the level of variation from the pre-assessment to the post-assessment. More specifically, 6 of the original 11 preservice teachers with the highest degree of variation on the pre-assessment remained in the group exhibiting the highest degree of variation on the post-assessment. By way of contrast, only 2 of the original 11 preservice teachers with the least variation in their pre-assessment responses remained in the group demonstrating the lowest degree of variation on the post-assessment.

A summary of the dominant modes of inferential reasoning on the post-assessment is shown in Table 4.2. Percentages in the table have been rounded.

Table 4.2

Dominant Levels of Reasoning on the Post-Assessment

Participants	Prestructural Number (%)	Unistructural Number (%)	Multistructural Number (%)	Relational Number (%)
All Preservice Teachers	2 (6%)	11 (33%)	15 (45%)	5 (15%)
Middle School Preservice Teacher	2 (13%)	8 (50%)	4 (25%)	2 (13%)
Secondary School Preservice Teachers	0 (0%)	3 (18%)	11 (65%)	3 (18%)

Upon completion of the statistics course, the dominant level of inferential reasoning for the cohort shifted from Unistructural to Multistructural. However, this shift only partially occured for the middle school preservice teachers, as the dominant level of reasoning for this set of participants remained at the Unistructural level. Figure 4.4 shows the combined results of the pre- and post-assessment for the middle and secondary mathematics preservice teachers as two different populations. The lighter dots represent the secondary preservice teachers, while the darker dots represent the middle school preservice teachers.

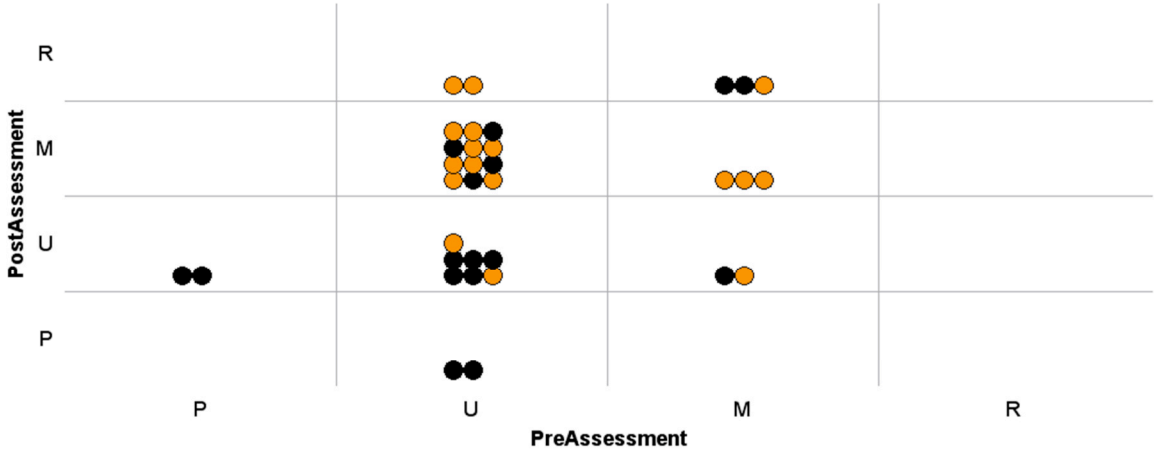


Figure 4.4. Middle and secondary mathematics preservice teachers' combined assessment results

From the data display in Figure 4.4, movement in inferential reasoning levels (both up and down) is observed, with most of the cohort beginning in the Unistructural level of reasoning and moving to the Multistructural level. The middle school preservice teachers tend to be grouped in the lower quadrant more than the secondary population and are the sole tenants of the Prestructural level. In addition, secondary preservice teachers reside in the Multistructural and Relational levels twice as often as their middle school peers.

The movement in each dominant level of reasoning is quantified in Table 4.3.

Table 4.3

Shifts in Levels of Inferential Reasoning, Pre- to Post-Assessment

Participants	Down One Level	Remained the Same	Up One Level	Up Two Levels
All Preservice Teachers	4 (12%)	10 (30%)	17 (52%)	2 (6%)
Middle School Preservice Teachers	3 (19%)	5 (31%)	8 (50%)	0 (0%)
Secondary School Preservice Teachers	1 (6%)	5 (29%)	9 (53%)	2 (12%)

Table 4.3 shows that almost one third of the cohort remained at the same dominant level of reasoning from the pre- to post-assessment, over half of the cohort moved up one or two levels in inferential reasoning, and a smaller percentage moved down one level. The shifts in dominant levels of reasoning were relatively similar between the middle and secondary preservice teachers. To illustrate the typical observed changes in dominant levels reasoning, I use five illustrative cases to provide a detailed accounting of these observed commonalities in responses. Two cases illustrate stability in inferential reasoning over time, two cases are representative of gains in inferential reasoning, and one case conveys a decline in inferential reasoning over time.

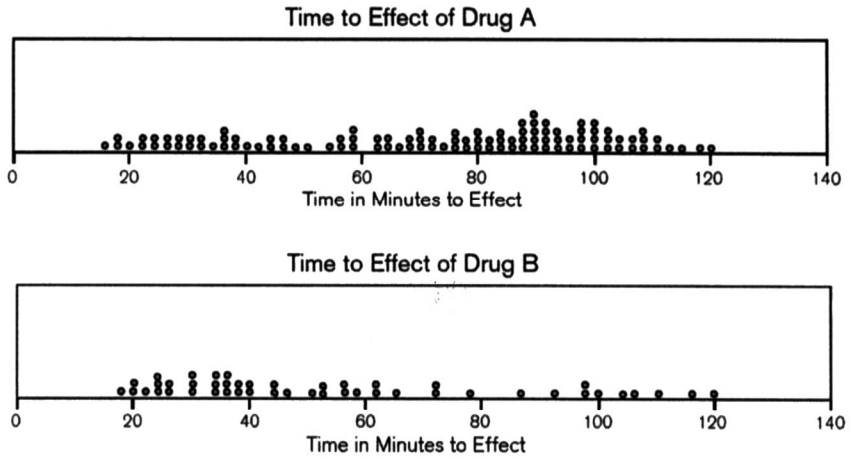
Stability in Inferential Reasoning Across Time

Stable Unistructural thinking. From the beginning of the statistics course through completion, 10 of the 33 preservice teachers demonstrated consistency in their dominant level of inferental reasoning across time. Most of these consistent reasoners, 7 of the 10, displayed a reasoning level of Unistructural on both the pre- and post-

assessments. Throughout the study, 21% the cohort consistantly displayed a dominant level of reasoning at the Unistructural level. Within this group, five members were middle school preservice teachers and two were secondary. Dave, a middle school preservice teacher, was typical in his stable use of Unistructural thinking across time. During the pre-assessment, Dave focused primarily on measures of center, such as mean or shifts in clusters of data, when reasoning on inferential tasks. One such example from the pre-assessment is shown in Figure 4.5.

Task 4: Which Treatment is More Effective?

Data are collected by two different pharmaceutical companies (Company A and Company B) on patients who suffer from migraine headaches. In both cases, the patients were told to take the medicine as soon as they experienced a headache and report how long it took to feel the relief effect of the medication. The results from each experiment are shown below.



Which medicine, Drug A or Drug B, would you recommend? Justify your choice below.

Drug B because the average reaction time was less than that of drug A.

Figure 4.5. Unistructural response to the Migraine Treatment task on the pre-assessment

In the preceding response, Dave focuses on the average response times of each medication and generates an inference based on which medication provides faster relief on average. Reasoning that attends to one aspect of the data is characterized as Unistructural. Other preservice teachers who also remained at the Unistructural level

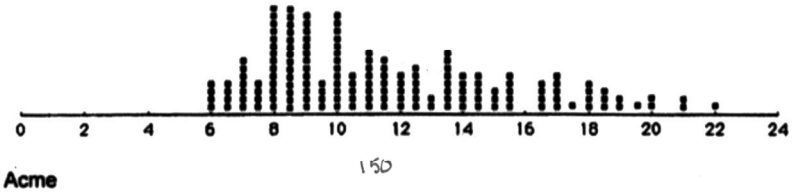
likewise focused primarily on measures of center and generated inferences based on differences beween measures of center.

During the mid-course assessment, Dave continued to focus on measures of center when generating inferences. In Figure 4.6, Dave reports that the difference in average response time is the key factor in considered to generate an inference despite a smaller sample. Since both samples are large, the difference in sample size is immaterial.

Task 4: Which Ambulance Service?

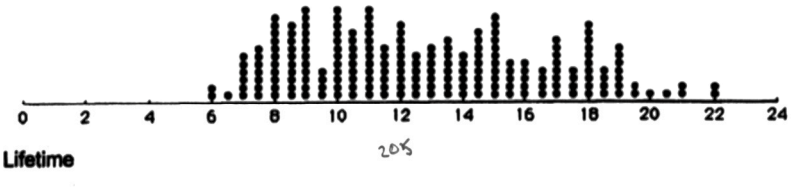
In St. Louis, the Clayton school district needs to select an ambulance service for emergencies that occur on school premises for the upcoming academic year. Two different ambulance companies provide service to the area: Acme and Lifetime.

Both companies provided the response times for emergency calls during the school year of 2009-2010 to other Clayton customers, as shown below. Acme provided data from 150 ambulance responses, and Lifetime provided data from 205 responses.



150

Acme



205

Lifetime

Based on the response times provided, which ambulance service would you recommend? Justify your choice below.

I would recommend Acme's service because their response times are quicker on average even though they reported less responses. I would expect the data to show a similar distribution if Acme had included 65 more responses.

Figure 4.6. Unistructural response to the Ambulance Service task on the midcourse assessment

While Dave's reasoning appears relatively unchanged from the pre-assessment, at times he begins to attend to other characteristics of the data distributions. One such example is in relation to a task that gauges the effectiveness of a speed trap designed to lower the speed of cars on a section of road. Initially, Dave's response appears to be consistent with prior arguments focused on measures of center, as depicted in Figure 4.7.

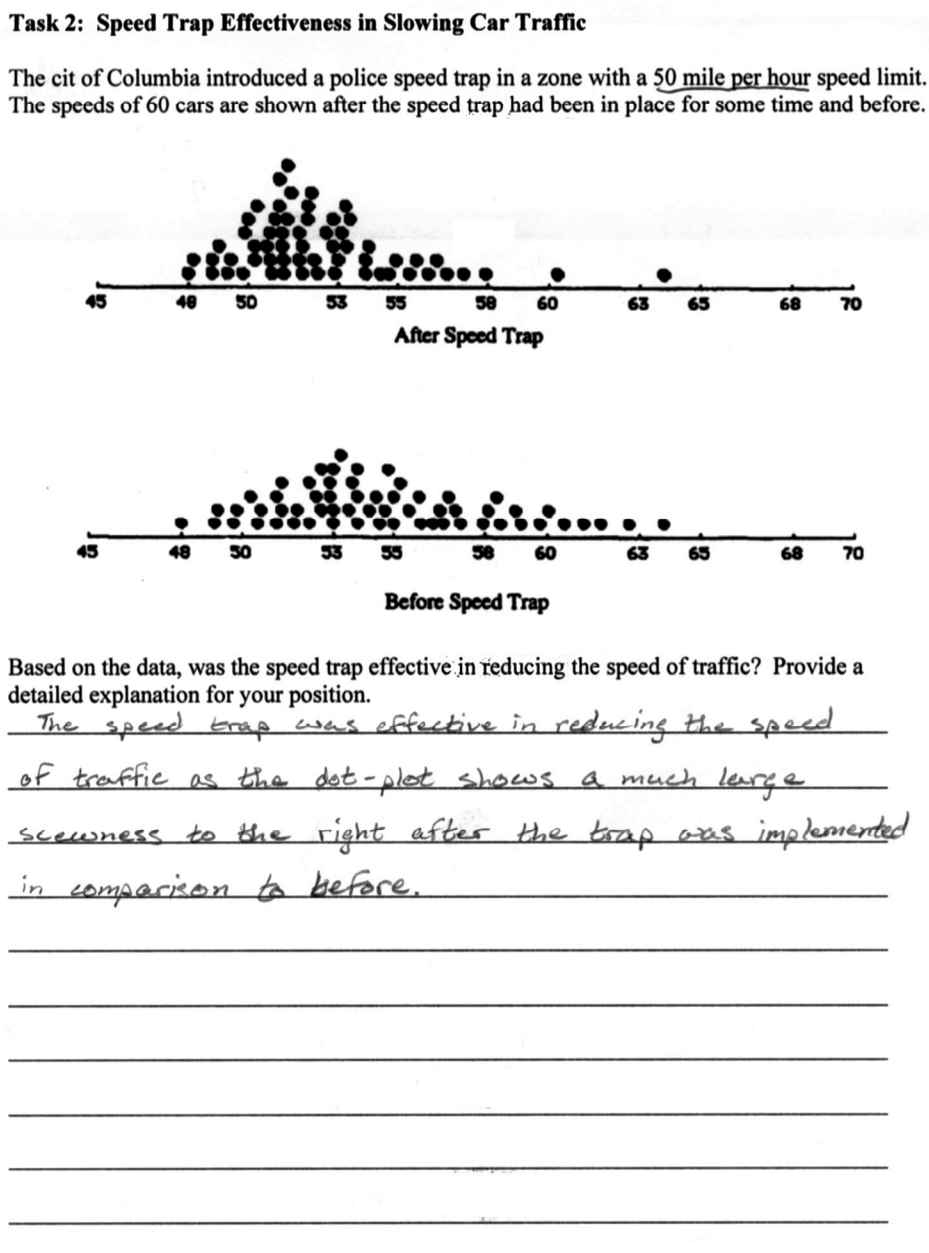


Figure 4.7. Unistructural to Multistructural reponse for the Speed Trap task on the midcourse assessment

Dave argues that the speed trap is effective as the data is skewed to the right after the speed trap was completed. When Dave was asked to elaborate on this written response, he explained that his primary consideration was the reduction in cars that sped over 58 miles per hour.

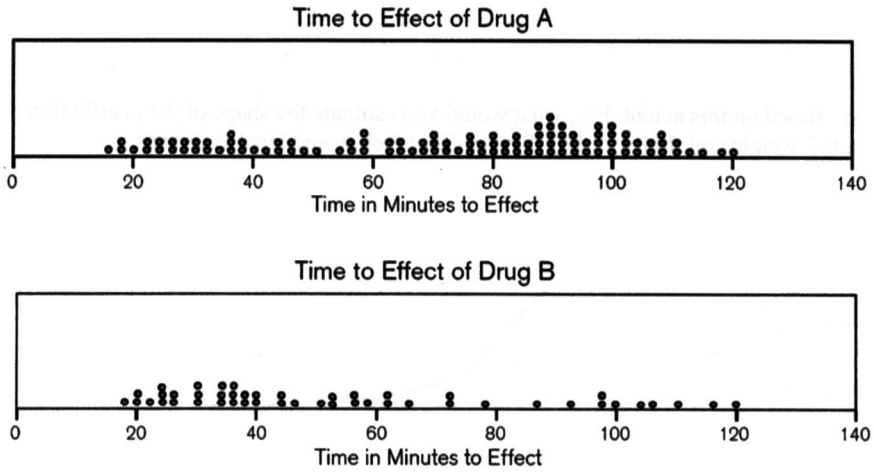
> Dave: It was effective in slowing traffic down. 'Cause it showed, because the dot plot is skewed a lot more to the right after, showing the after.
> MH: OK.
> Dave: Um, looking from, really looking from about 58 on there were, probably about 12 times that were larger, um, to the right of the, to the right of 58 before the speed trap and only 3 afterwards.
> MH: OK.
> Dave: So, it greatly, um, took down those top numbers.
> MH: Do you think that could have happened just by chance? 'Cause they just kind of picked 60 cars.
> Dave: Um.
> MH: So, do you think that could have been chance or to you, are you convinced?
> Dave: To me, it shows that the speed trap did work.
> MH: OK. So, for that main focus looking at 58 and above.
> Dave: Yeah.
> MH: Is there anything else that kind of supports it?
> Dave: Um, just that the average. I guess the median would be more towards between 50 and 55 where before it was between 53 and 55.
> MH: OK.
> Dave: Or 50 and 53, I guess after.
> MH: OK.

The exchange above demonstrates that Dave is able to consider other aspects of the data when reasoning beyond measures of center. In this particular case, Dave's main focus is on the tail-end of the data. Dave defines the goal of the speed trap to be a reduction in cars that travel in excess of 58 miles per hour. However, both in his written response and verbal explanation, Dave proffers a reduction in central tendancy as evidence of a successful speed trap. During the interview, Dave explains that the speed trap is effective because of the reduction in the number of cars that sped over 58 miles per hour. When asked for additional justification, Dave then mentions that the average

speed of cars has been reduced, similar to his written response. These two arguments, one based on Dave's intution about the context and one based on a formal measure of center, are not presented in a coordinated fashion in support of an inference that the speed trap was effective. Clearly, Dave has formulated a stance regarding the effectiveness of the speed trap based on data comparisons beyond changes in measures of center. From this exchange, either Dave is unable to reconcile his informal inference focused on the tail of the data distributions with measures of center or he believes that an argument based on measures of center is the correct response. Regardless, Dave continues to exhibit reasoning at the Unistructural level throughout the course. As shown in Figures 4.7 and 4.8, Dave's post-assessment responses to the Migraine Treatment and the Diet and Cholesterol tasks reveal how he limits his argumentation to measures of central tendancy.

Task 1: Which Treatment is More Effective?

Data is collected by two different pharmaceutical companies (Company A and Company B) on patients who suffer from migraine headaches. In both cases, the patients were told to take the medicine as soon as they experienced a headache and report how long it took until they felt the relief effect of the medication. The results from each experiment are shown below.



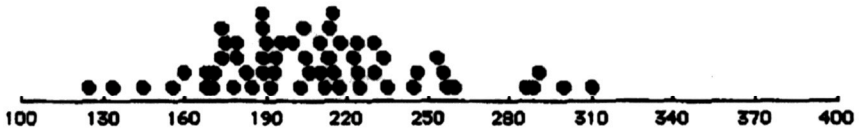
Which medicine, Drug A or Drug B, would you recommend? Justify your choice below.

I would recommend Drug B, because the mean time it took to see results were much less for drug B than they were for drug A.

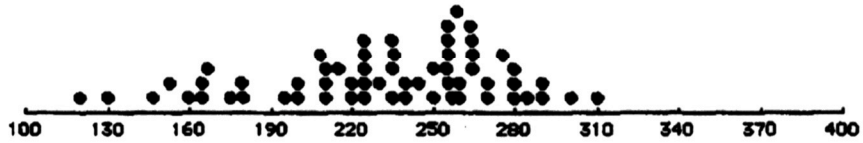
Figure 4.7. Unistructural response to the Migraine Treatment task on the post-assessment

Task 4: Diet and Cholesterol

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



After Program



Before Program

Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

The change in diet was effective because the mean cholesterol dropped significantly after the program.

Figure 4.8. Unistructural response to the Diet and Cholesterol task on the post-assessment

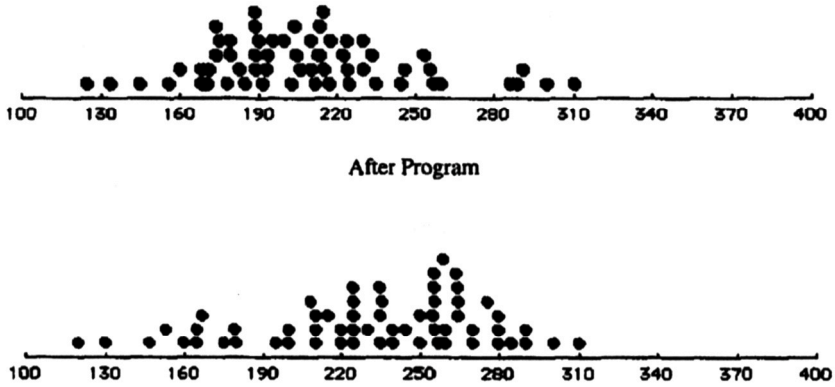
The portion of the cohort that remained unchanged in a dominant level of reasoning at the Unistructural level similarly tended to focus on measures of center when generating inferences and did not coordinate information related to other core statistical concepts such as range of the data distribution and variation in results.

Stable Multistructural Thinking. Another portion of the cohort remained unchanged in terms of their dominant level of reasoning at the Multistructural level. Specifically, eight preservice teachers initially exibited a dominant level of reasoning at the Multistructural level, and three of these remained at the same level upon completion of the statistics course, which represents appoximately 10% of the cohort. To illustrate the stable use of Multistructural inferential reasoning, consider the case of Cari, a secondary preservice teacher.

During the pre-assessment, Cari attended to measures of center, such as mean or shifts in clusters of data and another aspect of the data distribution when reasoning on inferential tasks. One such response from the pre-assessment is shown in Figure 4.9.

Task 6: Diet and Cholesterol

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

The two graphs show about the same range, so it is possible that the program doesn't lower cholesterol. But for the most part I would say the program is effective. Before the program, most participants had a level of approx 260. After the program, that number dropped to about 200. I think the program helps some, but not all participants.

Figure 4.9. Multistrucutral response to the Diet and Cholesterol task on the pre-assessment

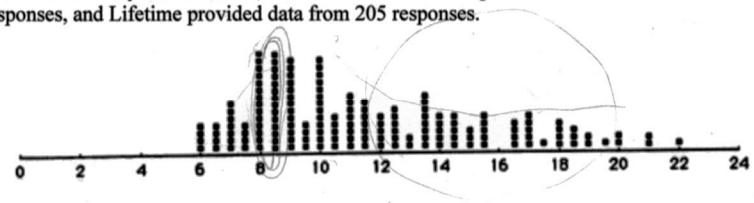
Cari begins the response by noting that the range in both the before and after data distributions are the same. She then discusses the reduction in average cholesterol levels from 260 to approximately 200. As a final step, Cari *coordinates* these two pieces of information by saying that the diet was effective in part, but did not help those who remained unchanged.

During the midcourse assessment and post-assessment, Cari generally provided Multistructural level responses and a few at the Relational level. Figure 4.9 shows a Multistructural response provided by Cari that coordinates the difference in centers and variation between two data distributions.

Task 4: Which Ambulance Service?

In St. Louis, the Clayton school district needs to select an ambulance service for emergencies that occur on school premises for the upcoming academic year. Two different ambulance companies provide service to the area: Acme and Lifetime.

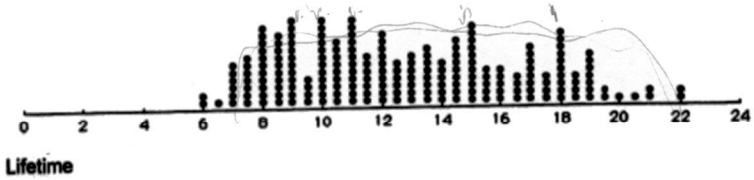
Both companies provided the response times for emergency calls during the school year of 2009-2010 to other Clayton customers, as shown below. Acme provided data from 150 ambulance responses, and Lifetime provided data from 205 responses.



Acme

↑ mode → about 8

Skewed right - most have lower times



Lifetime

→ mode not clear relatively flat, will center about 12.

Based on the response times provided, which ambulance service would you recommend? Justify your choice below.

I would recommend the Acme Service. Looking at the graphs, Acme's data is skewed right - this means that while some calls can take 18+ minutes, the majority of calls are around 8 or 10 minutes. The graph of Lifetime's data is more flat, and I approximate its center to be at 12 or 13. Also, the mode for Acme's data is fairly clear - about 8 min. But Lifetime's data doesn't have a mode - there are high points at 8½, 10, and 10½ minutes.

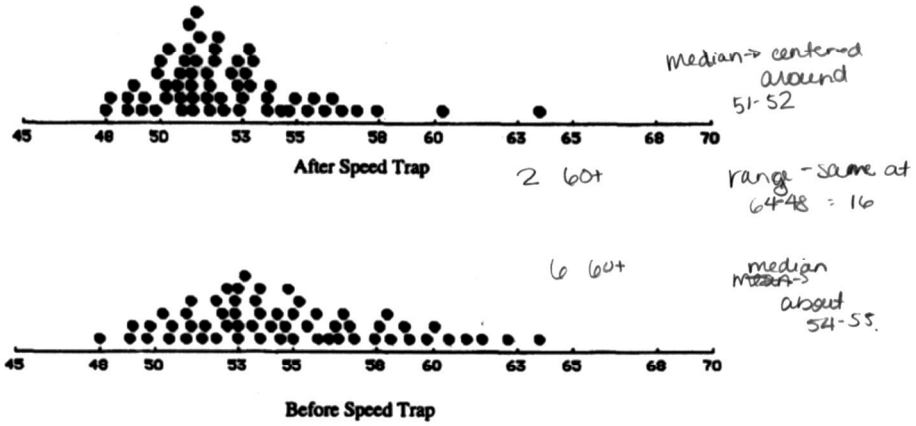
Figure 4.10. Multistrucutral response to the Ambulance Service task on the midcourse assessment

By taking in account differences between the data distributions in terms of measures of center and one additional global characteristic such as spread or variation in the data, Cari's response typifies others whose dominant level of reasoning remained at the Multistructural level. In this case, Cari is viewing the modes as indications of the variation within each data distribution. In particular, she argues that Lifetime's data distribution is relatively flat and has multiple modes, where ACME has a clear mode and is skewed to the right. This information further supports that ACME is a better choice than Lifetime, since the average response time is lower *and* there is less variation.

In addition to providing Multistructural responses on the midcourse assessment, Cari also provided one Relational response, which was common for others who remained at this level. An example of Cari's Relational response to the Speed Trap problem on the midcourse assessment is presented in Figure 4.11.

Task 2: Speed Trap Effectiveness in Slowing Car Traffic

The cit of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.



median → centered around 51-52

range - same at 64-48 = 16

median about 54-55.

Based on the data, was the speed trap effective in reducing the speed of traffic? Provide a detailed explanation for your position.

The speed trap was effective in reducing the speed of the traffic. The median speed of drivers after the speed trap is about 51 or 52, when before, the median was about 54 or 55. Also, there are a lot fewer drivers who went 58 mph or above after the speed trap (3) than before (13). This is enough for me to say that the speed trap was effective. A note on the range → the min and max speeds are the same in both cases, but I would think this is due to the drivers who always drive fast (no matter speed limits/traps, etc), and there will be drivers who tend to drive slower too. So the very high and very low speeds are to be expected.

Figure 4.11. Relational response to the Speed Trap task on the midcourse assessment

In contrast to responses at the Unistructural level, Cari is able to reconcile the reduced average speed of the cars with the fact that the range has not changed in regard to higher speeds. Therefore, some cars are still driving at high speeds despite the overall reduction in speeds. Similar to Dave, Cari chooses the speed of 58 miles per hour as excessive and determines that there are far less excessive speeders in the range of 58

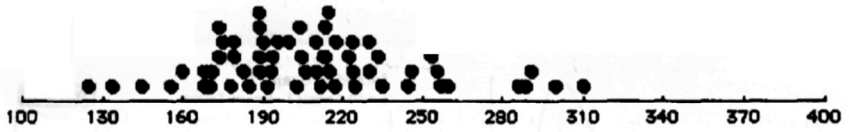
miles per hour and above, which further endorses the success of the speed trap and minimizes the fact that the range of the data has not changed. In contrast, those at the Unistructural level struggled to coordinate the small change in average speed with the reduction of excessive speeders in the tail of the data distribution. This response is designated as Relational due to all relevant aspects of the data distributions being identified and coordinated in order to provide a reasonable inference within the context of the problem. In support of this conclusion, Cari's verbalization of her reasoning process further endorses her coordination of all key aspects of this particular task:

> Cari: OK, I thought it was effective. The speed trap was effective because I looked at the center, the medians and I estimated that, um, after is about 51 or 52,
> MH: OK.
> Cari: And then before it was higher, about 54 or 55, and so, I figured, you know, a couple miles per hour, that's significant enough, um. Another thing I looked at was how high they went up, like the number of drivers who were about 58, so, after it was only 3, and then before I counted 13.
> MH: OK.
> Cari: And so, they definitely brought down the high, you know.
> MH: Really speeding, yes.
> Cari: The drivers who are really fast. Um, and then I just made a quick note on the range, because I noticed that the minimum and maximum were the same for both of them. And I just said for that, well, you're always going to have the drivers who just drive slow, and then you're always going to have at least a few that are just going to keep driving fast no matter what.
> MH: OK.
> Cari: So I figured that was to be expected, but overall it was still effective.

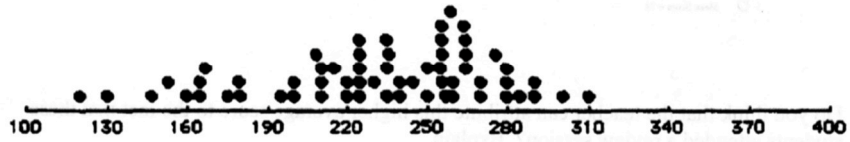
On the post-assessment, Cari continues to reason in a manner similar to the midcourse assessment at the Multistructural level. In the following example, Cari focuses on the change in measures of centers between the distributions and then attends to one global comparison between data distributions, the range of the data distributions.

Task 4: Diet and Cholesterol

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



After Program



Before Program

Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

I think that for most people, the new diet was effective. It looks like before the program, people tended to have levels between 220 and 260. After the program, that dropped to about 180 to 220. There are the few people who had extremely high or extremely low levels to begin with, and the program may not have affected them as much. But for the most part, I think the diet was effective.

Figure 4.12. A Multistructural response to the Diet and Cholesterol task on the post-assessment

The range in cholestorol values before and after the vegetarian diet remains unchanged in the task. Cari acknowledges the reduction in cholestrol levels for the majority of participants, but also incorporates the fact that some people did not benefit to provide a qualified inference based on the data. The response falls short of using all available information and data provided, but is qualitatively different than those at the

Unistructural level of inferential reasoning. Incorporating a global comparison in addition to considering changes in measures of center is typical of responses provided at the Multistructural level.

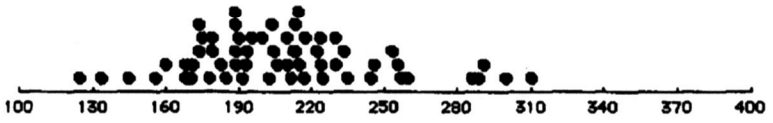
Growth in Inferential Reasoning Over Time

The most common change in dominant level of reasoning from the pre-assessment to the post-assessment was a one-level increase, with 17 of 33 (52%) members of the cohort moving up one level of inferential reasoning. Within this group, 12 total preservice teachers changed from the Unistructural reasoning level to the Multistructural level, 3 preservice teachers changed from the Multistructural level to Relational level, and an additional 2 changed from the Prestructural level to the Unistructural level.

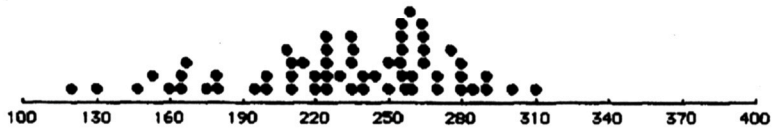
Unistructural to Multistructural. Because the largest trend in this study in terms of changes in inferential reasoning was the move from Unistructural to Multistructural reasoning, a set of responses from a middle school preservice teacher named Brandy will exemplify the changes documented in this group. In Figure 4.13, Brandy provides a response on a pre-assessment task that focuses on measures of center.

Task 6: Diet and Cholesterol

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



After Program



Before Program

Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

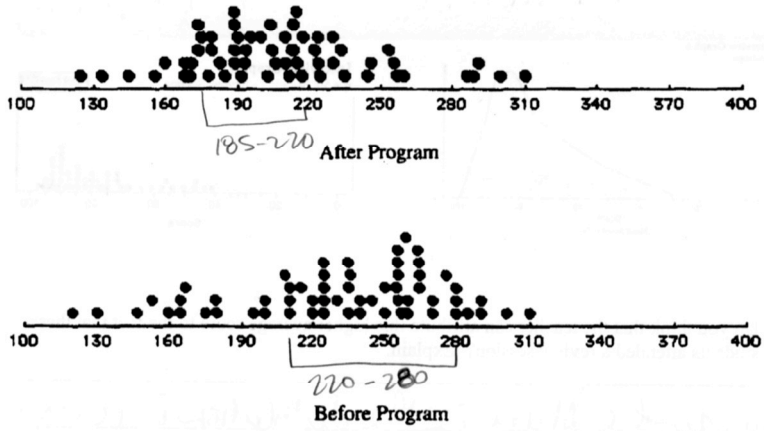
Yes, I would say that the change in diet was effective in lowering cholesterol, as most participants had levels between 100-280 before the program & 170-230 after the program.

Figure 4.13. A Unistructural response to the Diet and Cholesterol task on the pre-assessment

In this response, Brandy focuses on a change in the center value from before the vegetarian diet to after the diet and generates an inference that the diet was effective. However, during the post-assessment, Brady incorporates another aspect of the data to generate a Multistructural response.

Task 4: Diet and Cholesterol

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

I believe the results could have been achieved by chance more than the alternative diet. Although the peak cholesterol range drops from ~220-280 down to 185-220, the overall range of cholesterol levels of all participants is the same & has very similar distribution.

Figure 4.14. A Multstructural response to the Diet and Cholesterol task on the post-assessment

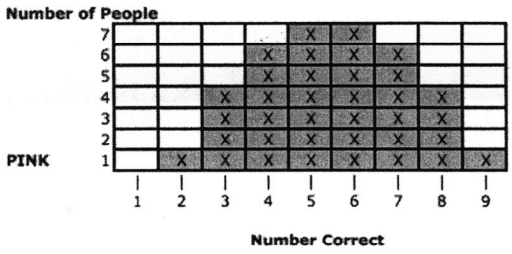
In this response, similar to the response provided by Cari, Brandy recognizes the fact that not all participants appear to benefit from the vegetarian diet, as the range has not changed in comparison to the original. Note that she concludes the change in the data could have occurred by chance, since the range and overall shape of the data look approximately the same. While the conclusion is not as well connected to the context of the problem as the one provided by Cari, Brandy does acknowledge two different aspects

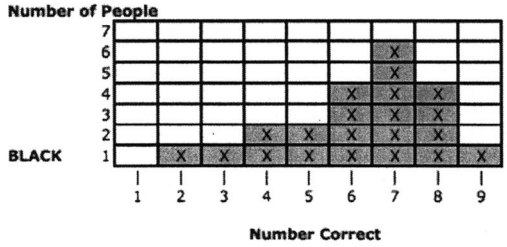
of the data, change in central tendancy and lack of change in range, and attempts to integrate them together in order to generate an inference. This particular response is typical of those in this group. In general, those who moved from Unistructural level responded to the pre-assessment in a similar manner to Dave and then responded similar to Cari on the post-assessment. While not all members of this particular group participated in the midcourse interview, Brandy's thinking exhibited ascension to a Multistructural level of reasoning at this point in the course.

Prestructural to Unistructural. Although only two middle school preservice teachers reasoned at the Prestructural level on the pre-assessment, both demonstrated Unistructural inferential reasoning by the end of the course. We will consider the case of a middle school preservice teacher named Ella, and her response to a complete population task given on the pre-assessment in which she struggles to proportionally reason.

Task 2: Which Class Did Better?

Two classes are competing on quick recall math facts. One class is called the "Pink" class, and the other the "Black". The two classes both complete a quiz, and the results are shown below.





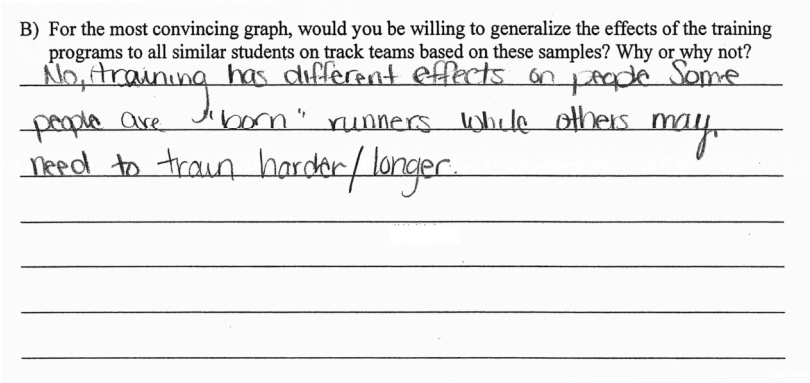
Which class did better? Please provide a complete explanation and any numerical information used in your rationale. The 2 classes are equal. The "Pink" class appears to have done better but there is just simply more students in the class Neither class had some do less than 2 wrong & only 1 person with 9 right.

Figure 4.15. A Prestructural response to the Class Scores task on the pre-assessment

Ella states that the two classes are equal in terms of their performance. However, by looking at the shape of the two populations, one can see that the Black class clearly has a higher percentage or proportion of studenst who scored at the levels of 7, 8 and 9

correct. While the *absolute* number of students scoring at these levels is the same, the *proportion* of students in the Pink class who scored at these higher values is less than that of the Black class. Ella states that the classes did equally well based on the range of the populations being the same. She also states that the Pink class appears to have performed better, perhaps because of the larger number of scores for 3, 4, 5 and 6 correct. However, the absolute frequency of scores for any specific value is irrelevant given that the populations are not of equal size. Because many of the tasks consisted of dot plot representations of data distributions, errors in proportional reasoning were observed several times in the pre-assessment and this was especially true for the preservice teachers classified at the Prestructural level, as well as for several whose dominant inferential reasoning level was Unistructural. Generally, preservice teachers who exhibited an inability to reason proportionally on the pre-assessment also struggeld to correctly apply proportional reasoning on the post-assessment.

In addition to a lack of proportional reasoning, those at the Prestructural level tended to answer tasks based on mere opinion statements that lacked any support based on the data or information provided. Ella's responses to the Training Programs tasks on the pre-assessment is provided as representative of this phenomenon.


B) For the most convincing graph, would you be willing to generalize the effects of the training programs to all similar students on track teams based on these samples? Why or why not?

No, training has different effects on people. Some people are "born" runners while others may need to train harder/longer.

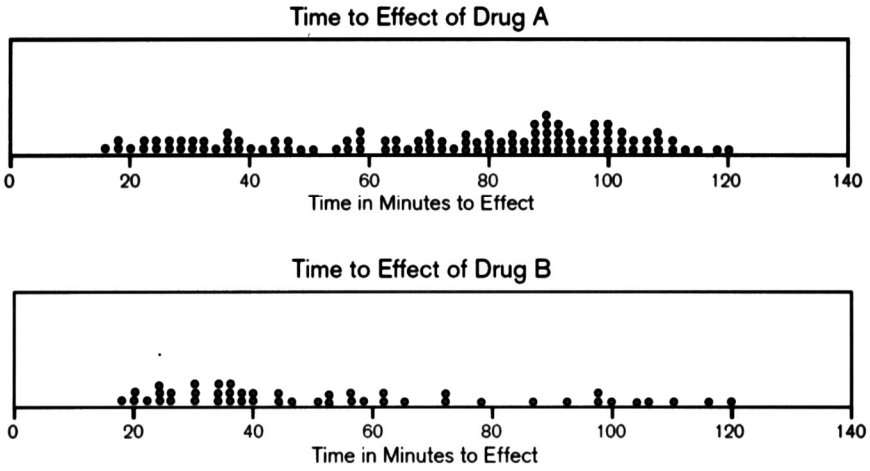
Figure 4.16. A Prestructural response to the Training Programs task on the pre-assessment

This response highlights that Ella is relying upon her personal opinions perhaps based on experiences or other factors in generating inferences. However, Ella appears to ignore the data provided in the task or the structural design of the experiment.

During the post-assessment, a substantial change was evident in Ella's responses, which became based on the data provided and consisted of one correct comparison to generate an inference. In the example provided in Figure 4.17, Ella attends to a comparison of *skewness.*

Task 1: Which Treatment is More Effective?

Data is collected by two different pharmaceutical companies (Company A and Company B) on patients who suffer from migraine headaches. In both cases, the patients were told to take the medicine as soon as they experienced a headache and report how long it took until they felt the relief effect of the medication. The results from each experiment are shown below.



Which medicine, Drug A or Drug B, would you recommend? Justify your choice below.

Drug B because the data is more distributed to the left where as Drug A the data is more to the right.

Figure 4.17. A Unistructural response to the Migraine Treatment task on the post-assessment

This particular task involves two data distributions that again have the same range similar to the pre-assessment task, Class Scores. However, on the post-assessment, Ella attends to the difference in *skewness* of the two data distributions, which indicates that one medicine will provide faster relief. Rather than specifically explaining that the skewness of the data distribution indicates faster or slower response times of the medicine, Ella simply notes the difference and skewess between the two distributions and

makes a claim. Ella continues throughout the post-assessment to offer justifications that are primarily based on the data provided in each task with one correct comparison.

During the midcourse assessment, an interesting exchange occurred regarding the Speed Trap task, revealing Ella's attention to *local* attributes of the data distributions versus *global* comparisons. Similar to the post-assessment, Ella provided a short explanation and claim regarding the effectiveness of the Speed Trap task in reducing the speed of cars shown in Figure 4.18.

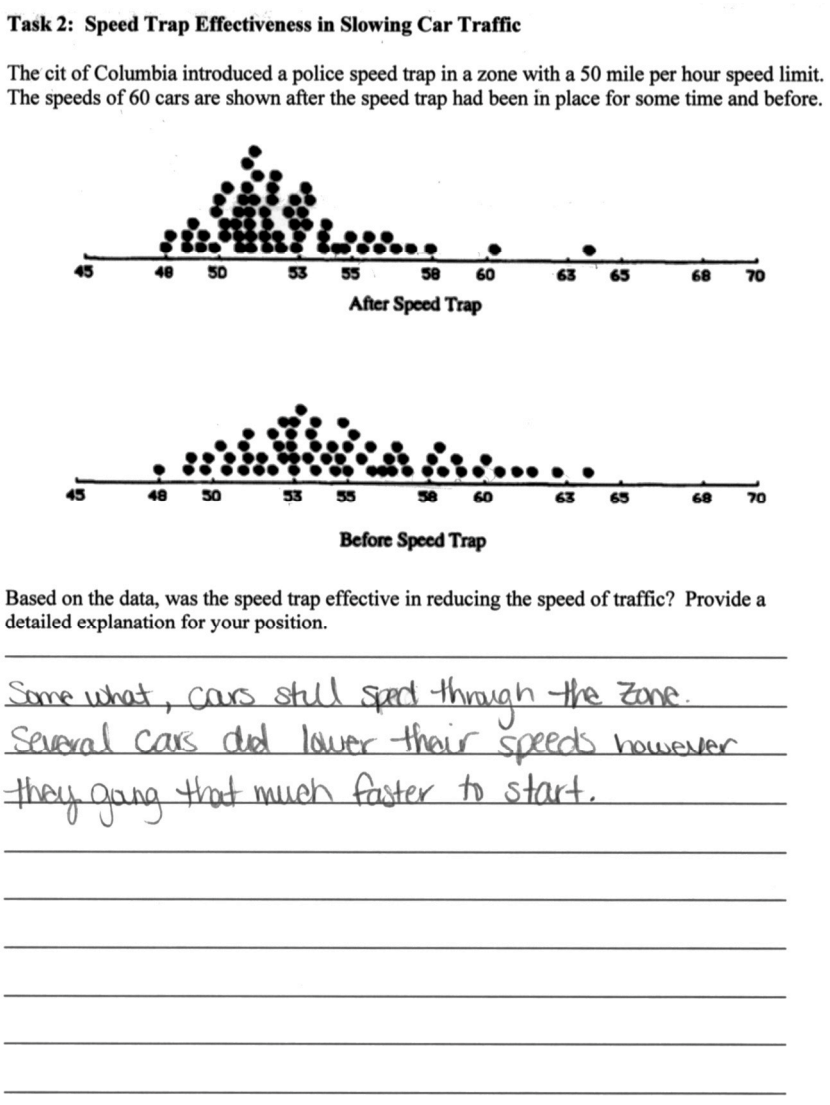


Figure 4.18. A Unistructural response to the Speed Trap task on the midcourse assessment

The written response focuses on the fact that the overall speed reduced only slightly after the speed trap was in place and that the average speed of the cars still exceed the desired speed limit of 50 miles per hour. However, when Ella verbally described her stance, the focus was not solely on the change in center, but also was concerned with outliers as evident in the following transcript:

> Ella: This one, I said No, because they weren't going that much over to start with. Like, they were only going three miles over. So, just backing off 3 doesn't mean it is a considerable change.
> MH: OK, so, yeah.
> Ella: Yeah, but effective, I said, "not", "somewhat", but not enough because there's still outliers. I mean they, still have these up here that weren't affected.
> MH: Yeah. OK. Alright, so say what you said the first, at the very beginning again.
> Ella: Oh, I said it was not effective because the majority was at 53, and that's not that much higher than 50, so they didn't really reduce that much of it.
> MH: Yea. So, you're saying because you would like to see, like these under 50, is that what you are saying? For it to be effective, or…?
> Ella: Or like. Yeah.
> MH: OK. So, they didn't really move that middle hump very much.
> Ella: Yeah. This clump right here. Or this clump right here really didn't move, cause there's still some there.
> MH: And, so that, that is one thing, and then you're looking at, uh, these ones that are way above, you're saying.
> Ella: The outliers, yeah.
> MH: You're saying didn't really move. There's still, there's still a group here.
> Ella: Yeah.
> MH: Even after. OK. Alright.

Ella deems the speed trap as only somewhat effective based on the fact that there are some cars still speeding above 58 miles per hour. Recall, Dave noticed that the number of cars over 58 miles per hour reduced from 12 to 3, which led him to infer the speed trap was successful, an inference shared by Cari. The distinction between Ella's inferential reasoning and that of Dave and Cari is that Ella focused on *local events* or *specific outliers*; whereas Dave and Cari viewed the tail the distribution as a region and

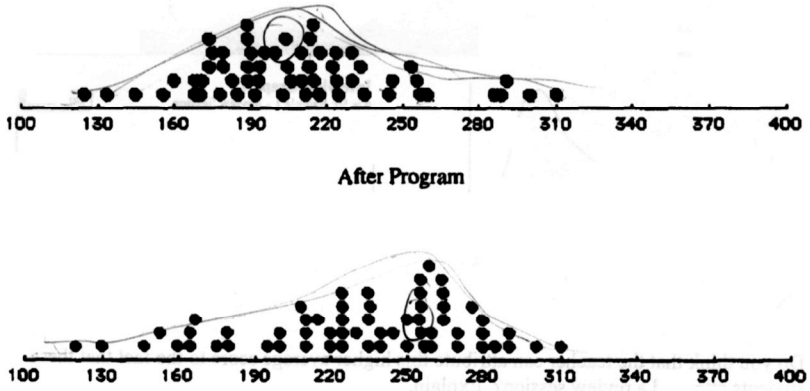
were able to make *global compaisons* about the changes in this portion of the data distribution rather than focusing on specific points within the tail.

Multistructural to Relational. Five preservice teachers progressed to the Relational level by the end of the course, two middle school and three secondary. Three of these changed from Multistructural to Relational from the pre-assessment to the post-assessment, and two from Unistructural to Relational. Rather than provide additional examples of Multistructural responses, the focus is on characteristics common to post-assessment tasks, as responses on the pre-assessment are similar in nature to those previously discussed.

Bretta, a preservice middle school teacher, represents the case of growth in her dominant level of inferential reasoning from Multistructural to Relational. Bretta had taken an introductory statistics course prior to this particular class, as was true of her middle school preservice teacher classmates. However, unlike her peers, Bretta's responses reflect her attention to *all data* provided in the task and a *coordination* of key statistical aspects with the context to provide a robust and complete response. As as an example of Relational inferential reasoning, consider Bretta's post-assessment response to the Cholesterol and Diet task in Figure 4.19.

Task 4: Diet and Cholesterol

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

Of course similar results could have been achieved by chance, but it looks like there has been a marked change in the medians of the two groups (the difference is somewhere around 50mg/dL?), which seems to say that the diet was effective. However, it seems unusual that some of the really high cholesterol people (presumably remained at high cholesterol levels – for some, there was no change. No one's cholesterol went up after the program, which means it can't hurt to go vegetarian. :) Overall, I think it was effective because the scores became concentrated around much lower values.

Figure 4.19. A Relational response to the Diet and Cholesterol task on the post-assessment

Bretta outlines the shape of the data distributions and circles where the centers may be, and subsequently provides an argument in support of the effectiveness of the diet based on an estimated change of 50 mg/dl in the average level of cholesterol, from before to after the diet. In addition to reporting the estimated change, Bretta believes the change is "marked", which supports the notion that the diet may significantly improve

cholesterol levels. However, the fact that some people's cholesterol levels did not decrease casts a measure of uncertainty on this claim. The response discussed above is qualitatively different than those at the Multistructural level because it *quantifies the change* in measures of center in order to provide justification for significance. Of those who responded to this task at a Multistructural level of reasoning, 60% failed to provide quantitative justification for significance. Moreover, in Relational inferential reasoning, significance is not rooted in personal opinion, but is rather based on a computation from the data presented. In addition, the small change in average speed lessened the support for significance in the minds of some preservice teachers.

One final illustrative example of the Relational level is a *formal* inferential reasoning response. The preservice teachers whose dominant level of reasoning was Relational were able to flexibly move between informal and formal approaches to tasks. The cohort's characterization of formal inferential reasoning on the post-assessment will be discussed in more detail in this chapter. In Figure 4.20, Bretta provides a formal response to the Speed Trap problem at a Relational level of reasoning.

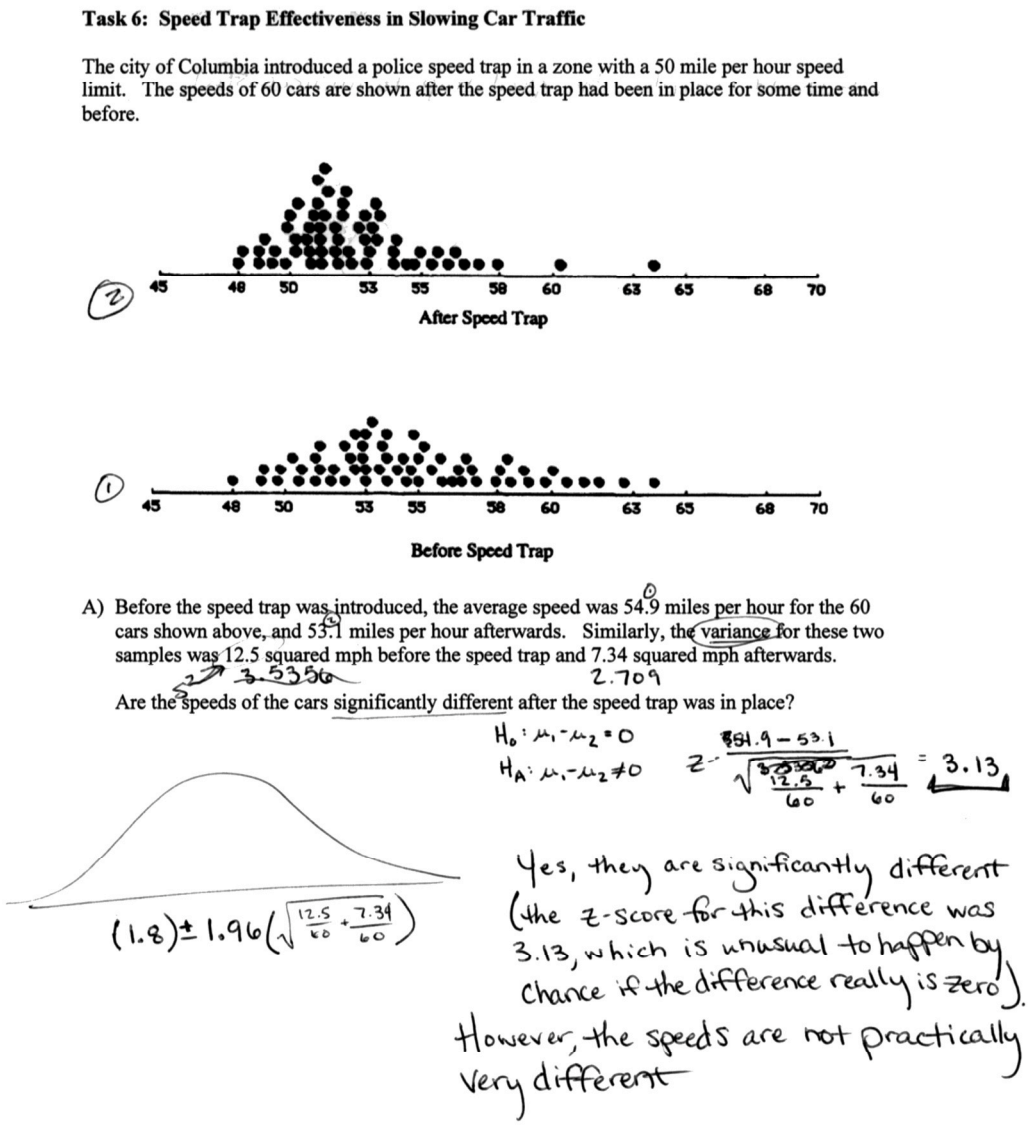


Figure 4.20. A Relational response to the Speed Trap task on the post-assessment

Bretta approaches the problem through a hypothesis test focused on the difference in means, as she was taught to do in class. She completes the computation correctly, interprets the results correctly, and provides a reasonable inference within the problem context. However, she notes that the speeds are not *practically* different, which certainly is true if the focus is on the average speeds (as opposed to reducing the amount of cars in excess of 58 miles per hour, as discussed previously). The difference of means

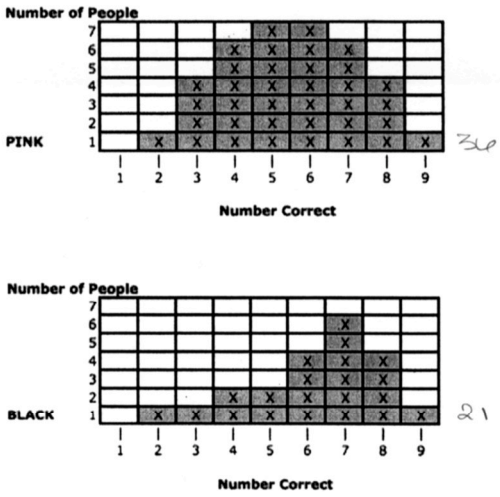
hypothesis test is not appropriate for this particular problem if the focus is on excessive speeders, but Bretta is not equipped with other formal methods. Her statement saying that the results did not match with her practical sense demonstrates that Bretta is attempting to make sense of the formal process as she would with informal approaches. Those who transitioned to the Relational level were able to demonstrate fluency in both formal and informal approaches to inferential tasks on the post-assessment.

Decline in Inferential Reasoning Over Time

Unistructural to Prestructural. While not sharing all of the characteristics of the previous group, the two middle school teachers whose informal inferential reasoning declined from Unistructural to Prestructural shared some similar characteristics. Both participants struggled with proportional reasoning from the onset of the statistics course. To illustrate preservice teachers' lack of proportional reasoning, consider Mary's response to the Class Scores task on the pre-assessment.

Task 2: Which Class Did Better?

Two classes are competing on quick recall math facts. One class is called the "Pink" class, and the other the "Black". The two classes both complete a quiz, and the results are shown below.



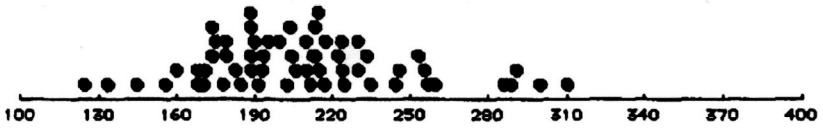
Which class did better? Please provide a complete explanation and any numerical information used in your rationale. Pink had more people but they got more correct

Figure 4.21. A Prestructural response to the Class Scores task on the pre-assessment

In contrast to Ella's earlier response focused on local versus global attributes, Mary is focusing on *absolute frequencies* within the data distributions. Neither Ella nor Mary are reasoning from a proportional perspective. However, on the pre-assessment, Mary is able to mainly focus on measures of center on other tasks and perform at the Unistructural level. One such example to the Cholesterol and Diet task is provided in Figure 4.22.

Task 6: Diet and Cholesterol

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



After Program

Before Program

Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

Change in diet is the reason for the change because the majority of the group moved down in cholesterol levels

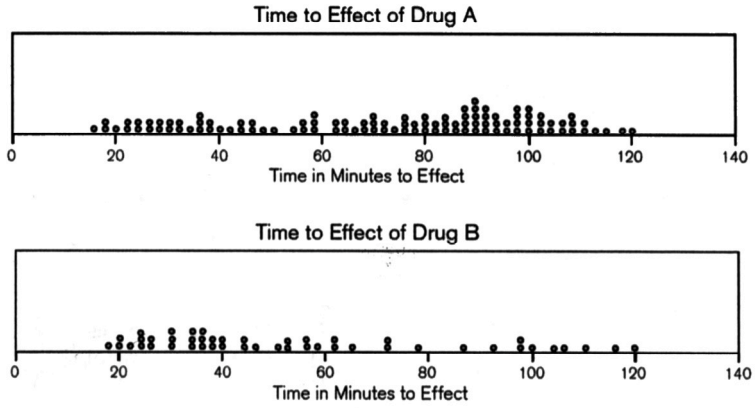
Figure 4.22. A Unistructural response to the Cholesterol and Diet task on the pre-assessment

Mary's response to the Cholesterol and Diet task focuses on a shift in the "majority" of data points or a change in central tendancy. However, during the post-assessment, Mary focuses on a frequency view of the data distributions and exhibits a Prestructural dominant level of thinking. A pair of tasks are shown that depict Mary

reasoning at a Unistructural level to Migraine Treatments task on the pre-asssessment and then at the Prestructural level on the post-assessment.

Task 4: Which Treatment is More Effective?

Data are collected by two different pharmaceutical companies (Company A and Company B) on patients who suffer from migraine headaches. In both cases, the patients were told to take the medicine as soon as they experienced a headache and report how long it took to feel the relief effect of the medication. The results from each experiment are shown below.



Which medicine, Drug A or Drug B, would you recommend? Justify your choice below.

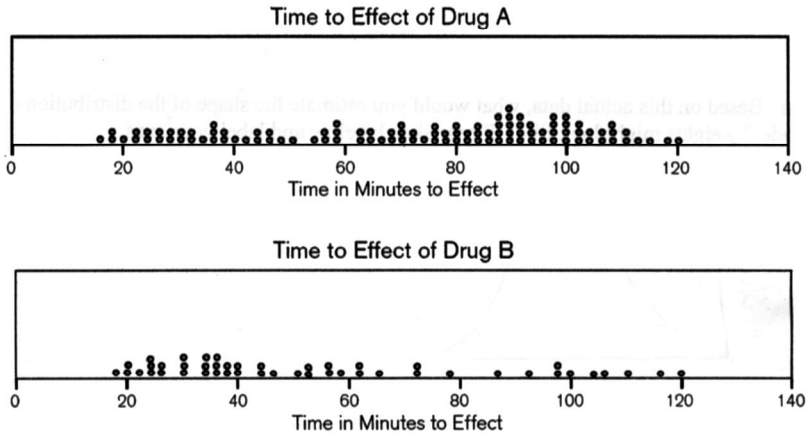
Drug B, although it had less participants, overall it's participants had less time than did that of Drug A

Figure 4.23. A Unistructural response to the Migraine Treatment task on the pre-assessment

In the above response, Mary ackowledges that the sample sizes of the two data distributions are not equal, and then states that the patients receiving Drug B experienced less time waiting for the medicine to be effective, seeming to support the use of proportional reasoning to make one correct data comparison to generate an inference.

Task 1: Which Treatment is More Effective?

Data is collected by two different pharmaceutical companies (Company A and Company B) on patients who suffer from migraine headaches. In both cases, the patients were told to take the medicine as soon as they experienced a headache and report how long it took until they felt the relief effect of the medication. The results from each experiment are shown below.



Which medicine, Drug A or Drug B, would you recommend? Justify your choice below.

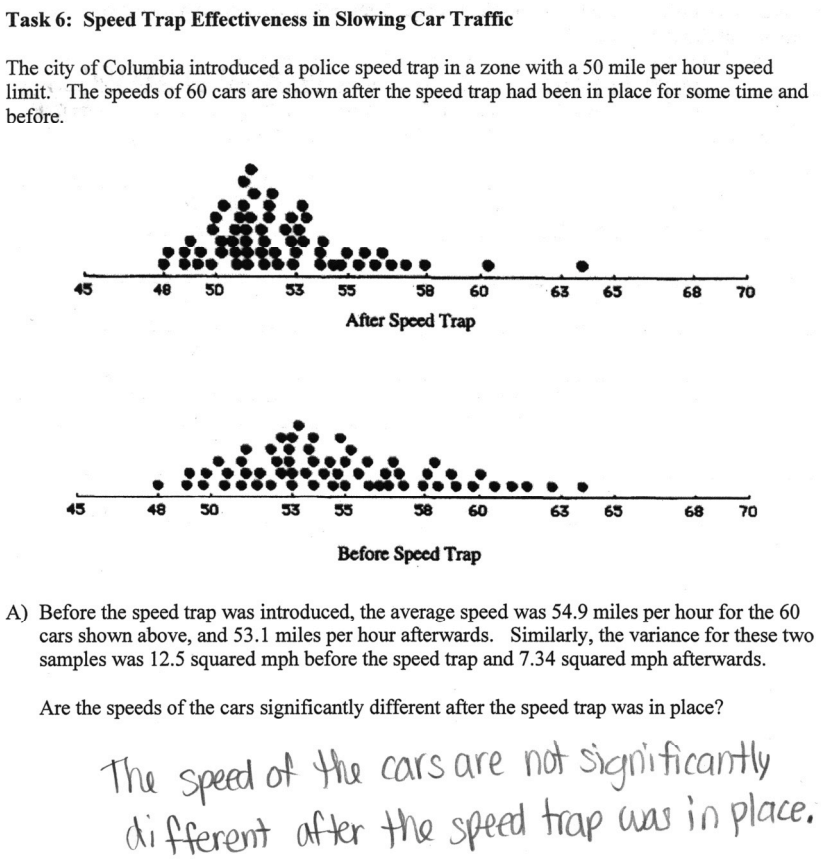
Drug A - Both drugs had about the same # of people positively respond before 40 minutes, but more people were tested for A

Figure 4.24. A Prestrurctural response to the Migraine Treatment task on the post-assessment

Mary ackowledges that the sample size of data in each distribution is different once again, but now states that the effect time of both medications is the same and that Drug A is a better choice due to the larger sample size. Note how Mary focuses on the frequencies of data points present in each sample from 0 to 40 and determines that the medicines provide an equivlanent time to relief of migraines. This task was generally one that most of the cohort addressed at either the Unistructural or Multistructural level

consistently. However, on the post-assessment, Mary provides a frequency focused response.

In addition to a focus on absolute (not relative) frequencies, Mary's post-assessment responses also were largely based on mere opinion. The tasks on the post-assessment asked the preservice teachers if the differences between data distributions were *significant*. This wording was used to signal that justification was desired either in the form of a hypothesis test, confidence interval or other statistical process. Mary often provided her opinion regarding whether the differences were significant rather than justifying significance based on changes or lack of changes in the data. An example is provided in Figure 4.25.



Task 6: Speed Trap Effectiveness in Slowing Car Traffic

The city of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.

A) Before the speed trap was introduced, the average speed was 54.9 miles per hour for the 60 cars shown above, and 53.1 miles per hour afterwards. Similarly, the variance for these two samples was 12.5 squared mph before the speed trap and 7.34 squared mph afterwards.

Are the speeds of the cars significantly different after the speed trap was in place?

The speed of the cars are not significantly different after the speed trap was in place.

Figure 4.25. A Prestrurctural response to the Speed Trap task on the post-assessment

While Mary's dominant level of reasoning was Prestructural in nature, similar responses to questions of significance were oberved for both Prestructural and Unistructural levels of reasoning. In Mary's case and her classmate who likewise performed at this level, most responses on the post-assessment either failed to incorporate proportional reasoning or were based on personal opinions instead of grounded in data.

Relation to Prior Statistics Coursework

The changes in dominant levels of reasoning over the course of the semester were unrelated to prior statistical coursework. The secondary preservice teachers possessed less prior statistical coursework, but began the course with higher levels of informal inferential reasoning and advanced their abilities at pace with middle school peers. The middle school preservice teachers tended to have more prior statistics coursework, but did not exhibit markedly higher levels of thinking than their secondary peers. In addition, the middle school preservice teachers did not demonstrate growth in their inferential reasoning that exceeded their secondary preservice teacher counterparts.

The secondary preservice teachers had less prior statistical coursework, but more prior mathematics classes such as calculus. The majority, 50%, of middle school preservice teachers' dominant level of reasoning on the post-assessment was Unistructural in nature versus 65% at the Multistructural level for secondary. In addition, the secondary preservice teachers experienced slightly higher positive changes in dominant levels of inferential reasoning than their middle school peers. Aside from these differences, responses to tasks possessed commonalities across the entire cohort, as did the patterns of change from one level to another and for those who did not change.

Characterization of Formal Inferential Reasoning

The findings specific to the cohort's inferential reasoning, which includes both *informal* and *formal* approaches on assessment tasks, were included in the characterization of the cohort's change in inferential reasoning. The characterization specific to the of the cohort's formal inferential reasoning is limited in nature due to timing of teaching formal methods in the statistics course. In particular, because formal inferential methods such as hypothesis testing and confidence intervals were taught during the last month of the course, preservice teachers were equipped with formal methods only for the post-assessment. Therefore, the cohort's responses to inferential tasks on the pre- and midcourse assessments were exclusively informal in nature.

Nevertheless, a characterization of the cohort's formal inferential reasoning can be characterized based on responses to the final three items on the post-assessment. These items, Speed Trap, Hiring Discrimination, and Ambulance Service, clearly signaled that a formal approach to generating inferences was desired (See Appendix C). Accordingly, most of the cohort attempted to generate an inference through formal approaches. However, when preservice teachers neglected how to implement a procedure or interpret a numerical result, occasionally they would default to informal reasoning and address the task similar to others in the assessment. In cases where a complete informal response was offered, the task was coded as being addressed informally.

Task responses were coded in the same manner as the informal inferential task responses and received either a Prestructural (P), Unistructural (U), Multistructural (M), or Relational (R) designation according to the SOLO taxonomy descriptions in Table 3.6.

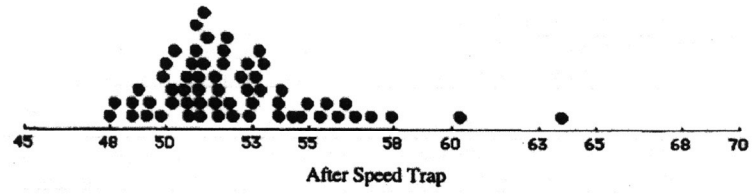
Variation in the responses given by the preservice teachers was observed both at the class level for specific tasks and at the individual level across tasks. However, because the number of tasks addressed formally by the each preservice teacher is small, summary results at the *individual* level across tasks are not provided.

Variation in Formal Inferential Reasoning Responses at the Class Level. The levels of formal inferential reasoning for the three post-assessment tasks are similar to other tasks on the post-assessment. The mean and standard deviation for three tasks addressed primarily through formal methods are: 1.67 and 1.08 for the Speed Trap task, 1.42 and 0.96 Hiring Discrimination, and 1.91 and 0.95 for the Ambulance Service task. Therefore, inferential reasoning response levels to all three tasks were similar to the overall mean on the post-assessment of 1.7 and fell between Unistructural and Multistructural levels. In addition, the variation in responses was similar to those previous discussed with standard deviation values near 1.0, which elicited responses across the full spectrum of levels of inferential reasoning. To illustrate the typical variation of formal inferential reasoning responses to a given task, archetypal responses of each level of coding are described for the Speed Trap on the post-assessment.

An example of a Unistructural (U) response to the Speed Trap task is shown in Figure 4.26. The preservice teacher identifies an appropriate hypothesis test for the task, but is unable to appropriately populate the values or draw an inference.

Task 6: Speed Trap Effectiveness in Slowing Car Traffic

The city of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.



After Speed Trap

Before Speed Trap

A) Before the speed trap was introduced, the average speed was 54.9 miles per hour for the 60 cars shown above, and 53.1 miles per hour afterwards. Similarly, the variance for these two samples was 12.5 squared mph before the speed trap and 7.34 squared mph afterwards.

Are the speeds of the cars significantly different after the speed trap was in place?

$$\frac{54.9-53.1}{\sqrt{\frac{(12.53)}{60}+\frac{(7.34)^2}{60}}}=.96185$$

Yes, I would say they are significantly different

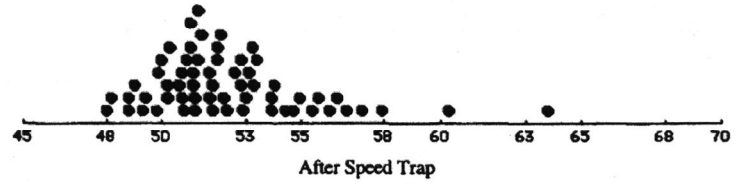
Figure 4.26. A Unistrurctural response to the Speed Trap task on the post-assessment

The preservice teacher chooses a *difference in means* hypothesis test, a reasonable selection, but incorrectly squares the variance values. In addition, the interpretation of the result is also incorrect, as a *z*-score of .96 is not indicative of significance.

Next, an archetypal Multistructural response to this same task is provided with a reasonable selection of a hypothesis test, correct population of values in the hypothesis test, but an unreasonable interpretation or inference. The Multistructural level response in Figure 4.27 entails a reasonable approach to addressing the task through formal methods

Task 6: Speed Trap Effectiveness in Slowing Car Traffic

The city of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.



After Speed Trap

Before Speed Trap

A) Before the speed trap was introduced, the average speed was 54.9 miles per hour for the 60 cars shown above, and 53.1 miles per hour afterwards. Similarly, the variance for these two samples was 12.5 squared mph before the speed trap and 7.34 squared mph afterwards.

Are the speeds of the cars significantly different after the speed trap was in place?

$H_0: \mu = 50$
$H_A: \mu \neq 50$

$$z = \frac{(54.9 - 53.1) - 0}{\sqrt{\frac{12.5}{60} + \frac{7.34}{60}}} = \frac{1.8}{.575} = 3.13$$

.2083 + .1223

No

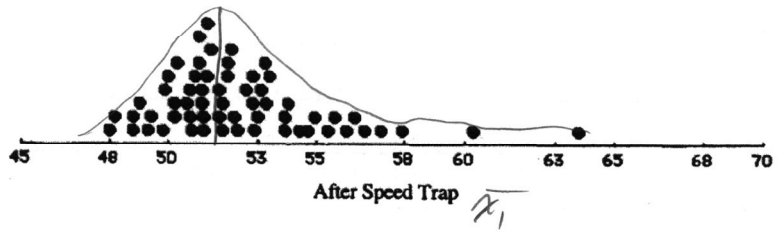
$z = 3.13$

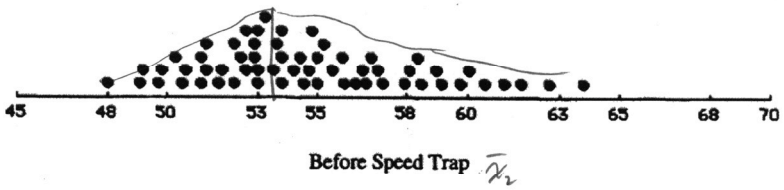
Figure 4.27. A Multistructural response to the Speed Trap task on the post-assessment

The computation produces a *z*-score test statistics of 3.13, which is a significant result for a two-tailed hypothesis test at the 90% confidence level. Accordingly, this interpretation of the hypothesis test and associated inference are not reasonable. Lastly, an archetypal response at the Relational level is provided which integrates all components of a formal approach and produces a sound inference.

Task 6: Speed Trap Effectiveness in Slowing Car Traffic

The city of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.



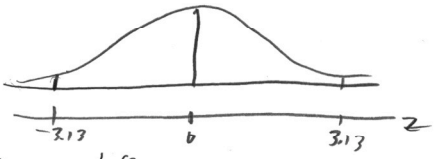


A) Before the speed trap was introduced, the average speed was 54.9 miles per hour for the 60 cars shown above, and 53.1 miles per hour afterwards. Similarly, the variance for these two samples was 12.5 squared mph before the speed trap and 7.34 squared mph afterwards.

Are the speeds of the cars significantly different after the speed trap was in place?

$H_0: \bar{x}_1 = \bar{x}_2$

$H_a: \bar{x}_1 \neq \bar{x}_2$



$$z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{7.34}{60} + \frac{12.5}{60}}} = \frac{53.1 - 54.9}{\sqrt{.331}} = \frac{-1.8}{.575} = -3.13$$

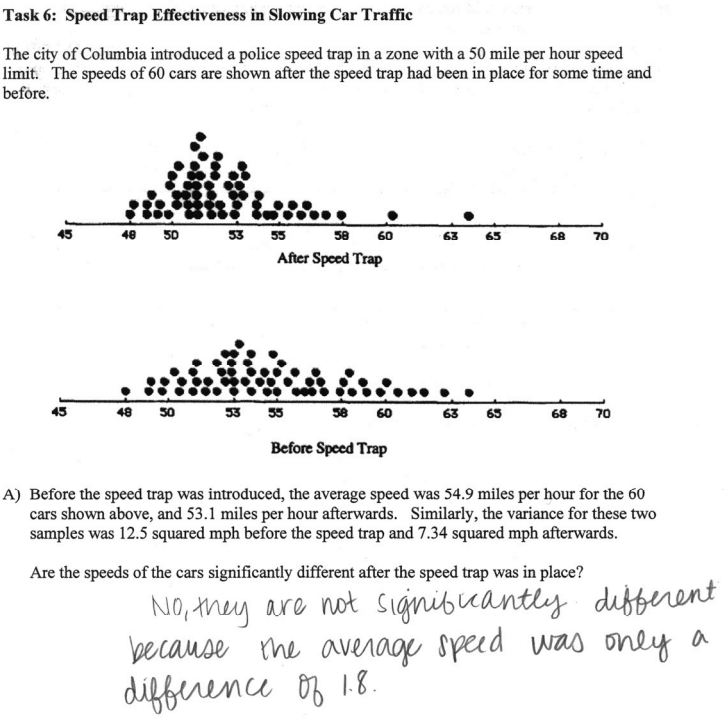
@ 90% CI, CV = ±1.645

∴ statistically significant difference.

Figure 4.28. A Relational response to the Speed Trap task on the post-assessment

The response illustrates a reasonable formal approach, correct population of variables, and a logical inference based on the results of the approach. The preservice teacher demonstrates additional knowledge of the procedure by drawing the probability distribution associated with the *z*-score and the fact that a two-tailed hypothesis test was selected.

Common Problematic Features of Formal Inferential Reasoning. Two common difficulties emerged on the three formal inferential tasks: (a) interpreting significance for 28% of responses, and (b) inputting the incorrect values into the appropriate hypothesis test for 32% of responses. All three tasks entailed a question relating to if the difference between two data distributions was significant in nature given a specific context. During the statistics course, significance levels were explicitly stated for the preservice teachers during lecture and in homework exercises. While formal approaches to these tasks reported large *z*-values, many preservice teachers claimed that the results did not support a significant difference similar to the response provided in Figure 4.27. Many also noted that the lack of significance made sense because the absolute difference in means was quite small in the case of the Speed Trap task and did not attempt a formal approach as shown in Figure 4.30.



Task 6: Speed Trap Effectiveness in Slowing Car Traffic

The city of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.

45 48 50 53 55 58 60 63 65 68 70

After Speed Trap

45 48 50 53 55 58 60 63 65 68 70

Before Speed Trap

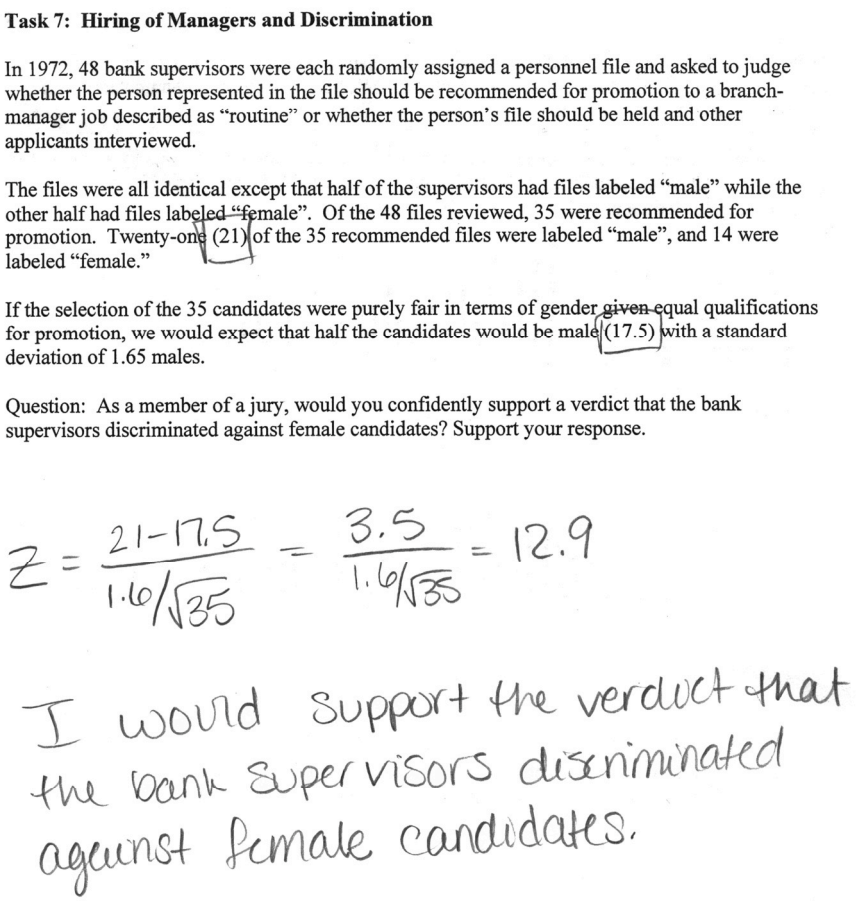
A) Before the speed trap was introduced, the average speed was 54.9 miles per hour for the 60 cars shown above, and 53.1 miles per hour afterwards. Similarly, the variance for these two samples was 12.5 squared mph before the speed trap and 7.34 squared mph afterwards.

Are the speeds of the cars significantly different after the speed trap was in place?

No, they are not significantly different because the average speed was only a difference of 1.8.

Figure 4.29. A Prestructural response to the Speed Trap task on the post-assessment

The second problematic characteristic was choosing an inappropriate hypothesis test or populating an appropriate choice with incorrect values. Preservice teachers generally had difficulty choosing a correct hypothesis test for a task with a binomial distribution. If they were able to identify the correct test, populating the values correctly posed challenges for many of the preservice teachers. Often, they would square the variance, leave D_0 in the equation, and/or put incorrect values in for the sample sizes. An example of an incorrect choice of formal approach to the Hiring Discrimination task on the post-assessment is provided in Figure 4.30.



Task 7: Hiring of Managers and Discrimination

In 1972, 48 bank supervisors were each randomly assigned a personnel file and asked to judge whether the person represented in the file should be recommended for promotion to a branch-manager job described as "routine" or whether the person's file should be held and other applicants interviewed.

The files were all identical except that half of the supervisors had files labeled "male" while the other half had files labeled "female". Of the 48 files reviewed, 35 were recommended for promotion. Twenty-one (21) of the 35 recommended files were labeled "male", and 14 were labeled "female."

If the selection of the 35 candidates were purely fair in terms of gender given equal qualifications for promotion, we would expect that half the candidates would be male (17.5) with a standard deviation of 1.65 males.

Question: As a member of a jury, would you confidently support a verdict that the bank supervisors discriminated against female candidates? Support your response.

$$Z = \frac{21-17.5}{1.6/\sqrt{35}} = \frac{3.5}{1.6/\sqrt{35}} = 12.9$$

I would support the verdict that the bank supervisors discriminated against female candidates.

Figure 4.30. A Prestructural response to the Hiring Discrimination task on the post-assessment

The response shown not only illustrates an incorrect selection of formal test, but a nonsensical interpretation of values relating to variables. Preservice teachers who

implemented confidence interval approaches to the last three tasks tended to avoid issues related to choice of formal test and reason more effectively.

In summary, the formal inferential reasoning characterization of the cohort specific to tasks on the post-assessment paralleled the overall inferential reasoning results. The cohort tended to reason between the Unistructural and Multistructural levels, but closer to Multistructural. Two problematic areas related to interpreting the statistical meaning of significance, choice of appropriate formal methods, and populating values incorrectly.

Association of Informal and Formal Inferential Reasoning

A comparison of the levels of formal inferential reasoning on the final three tasks of the post-assessment was made with the dominant level of informal inferential reasoning on the post-assessment. If the reasoning level of the formal task was the same as the dominant informal inferential reasoning level for a specific preservice teacher, the task responses was coded as in alignment or *concordant*. If the levels of reasoning did not match, the task was coded as *discordant*. Through this process, 80% of the 49 formal inferential reasoning task responses were found to be concordant with the preservice teacher's dominant informal reasoning level on the post-assessment. Multistructural and Relational formal inferential level responses were both coded as concordant to Multistructural dominant levels of informal inferential reasoning, since there were no dominant Relational levels of informal inferential reasoning on the post-assessment.

Given the high percentage of agreement in inferential reasoning levels between formal and informal approaches for a specific preservice teacher on the post-assessment, it appears that a relatively strong relationship exists between a preservice teacher's ability

to inferentially reason both informally and formally. In addition, the preservice teachers tended to reason at the similar levels for both formal and informal approaches on the post-assessment. Figure 4.31 shows the number of formal inferential reasoning responses provided at each level on the post-assessment, by area of content certification.

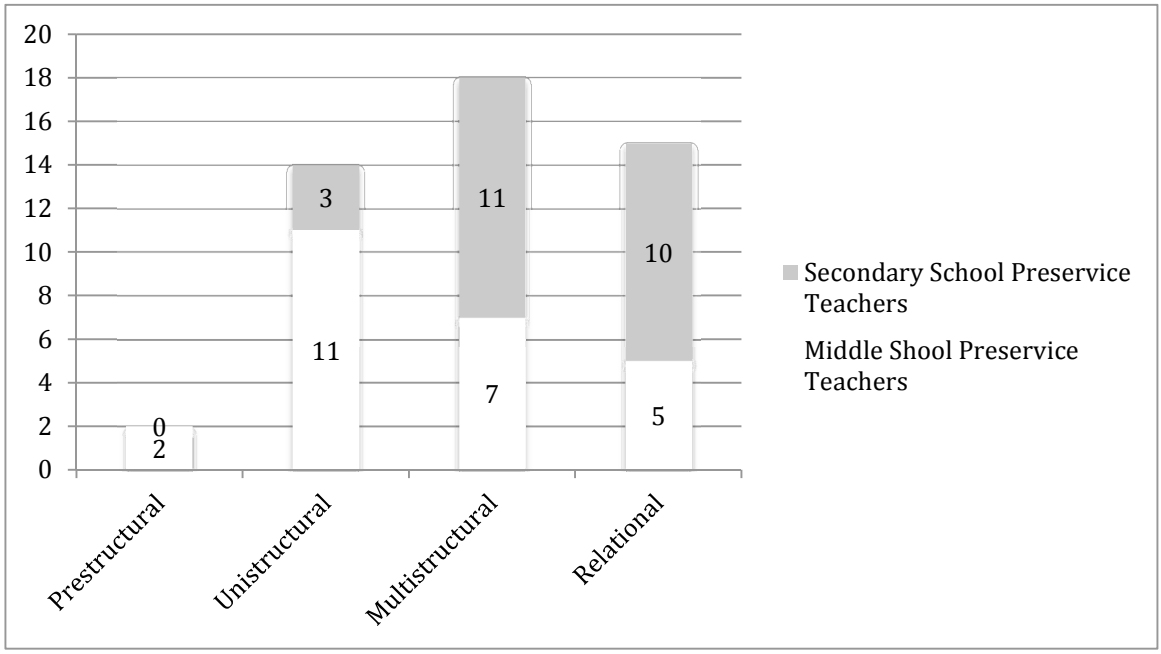


Figure 4.31. Levels of formal inferetial reasoning responses provided on the post-assessment tasks by area of certification

Given that the responses to tasks on the post-assessment were predominantly informal in nature, 182 informal responses compared to 49 formal responses, the alignment of formal response levels to dominant levels of informal inferential reasoning is remarkable. For example, middle school preservice teachers tended to provide the bulk of lower level responses at the Prestructural and Unistructural levels similar to their dominant levels of reasoning on the post-assessment. Similarly, secondary school preservice teachers tended to provide either Multistructural and Relational formal responses to tasks aligning with their primary dominant modes of inferential reasoning.

Therefore, it appears that a relationship between informal and formal inferential reasoning does indeed exist.

Opportunity to Learn

The cohort's opportunity to learn inferential reasoning is achieved through an analysis of the 375 tasks used within the statistics course. In accordance with the conceptual framework (See Figure 1.1), each task was coded in relation to core and aggregate statistical concepts and use of formal statistical methods such as confidence intervals or hypothesis tests. As previously stated, tasks could be assigned multiple codes, for content areas and formal methods. Therefore, the total number of content and methods codes assigned, 696, is larger than the number of total tasks, 375. On average, each task received 1.89 content and methods codes. In addition to coding for content and methods, each task was coded to one or more of the strands of mathematical proficiency as outlined in *Adding it up* (Kilpatrick, Swafford & Findell, 2001). Again, a single task may encompass multiple strands of mathematical proficiency, which resulted in 587 total codes for the same 375 tasks, an average of 1.56 codes per task. Lastly, each task was coded to determine if the statistical questions were contextualized. In total, 254 (or 68%) of the tasks were embedded in a context.

The course was organized into four sections: (a) descriptive statistics, (b) probability and probability distributions, (c) sampling and estimation, and (d) confidence intervals and hypothesis testing. Table 4.4 displays the proportion of the statistics dedicated to each section in terms of percentage of overall class time and the percentage of total tasks.

Table 4.4

Proportion of the Course Dedicated to Each Section of Content

	Tasks	% of Tasks	Weeks of Class	% of Class
Descriptive Statistics	136	20%	4	27%
Probability and Probability Distributions	263	38%	6.5	43%
Sampling and Estimation	103	15%	1.5	10%
Confidence Intervals and Hypothesis Testing	194	28%	3	20%

The course content was presented in the sequence provided by the required textbook, *Introduction to Probability and Statistics* by Mendenhall, Beaver & Beaver (2005) with supplementation from the *Quantitative Literacy Series* (Mrdulla et al., 1995). Specifically, the content began with the first chapter of the textbook related to describing data with graphs, and ended with the ninth chapter focused on formal hypothesis testing for the difference between means and proportions. Within the section of descriptive statistics, the content areas of graphing, measures of center and variability, and describing bivariate data are included. The second content section of probability and probability distributions includes finding probabilities of simple events, combinations, permutations, Bayes' Theorem, and discrete random variables. In addition, the Poisson and Binomial discrete distributions and Normal continuous distribution with associated approximations are discussed in this section. The next section relates to sampling, the Central Limit Theorem, experimental design including random sampling, sampling distributions, and

sampling variability. In the fourth and final section, confidence intervals and large-sample hypothesis testing are approached through traditional (algorithmic) approaches.

The tasks for the statistics course predominantly originate from the course textbook. In addition to exercises to be completed with pencil, paper, charts and calculators, computer simulation activities were also assigned from the textbook. Depending on directions given by the course instructor, the cohort completed these computer simulations either by running an applet provided with the textbook on a CD-ROM or by executing short Minitab programs in computer labs; the instructor provided the programming commands associated with writing and running the Minitab programs. Activities from the supplemental materials and were assigned and completed during class session by pairs of preservice teachers. During the class periods, material was introduced and explained by the instructor that differed from tasks found with the course materials. The tasks introduced by the instructor for instructional purposes were documented as part of this analysis as were the questions posed by the instructor verbally. Lastly, the tasks on the four exams were collected and coded. Table 4.5 summarizes the percentage of tasks and the associated source. In addition, the percentage of statistical content and method codes originating from each source is provided. Similarly, the percentage of strands of mathematical proficiency codes related to each source is shown.

Table 4.5

Proportion of the Tasks and Codes Generated from Each Source

	Percentage of tasks	Percentage of Content Codes	Percentage of Proficiency Codes
Textbook	54%	54%	51%
Computer Simulations	4%	6%	6%
Instructor	16%	18%	18%
Quantitative Literacy Series	13%	11%	10%
Exams	13%	11%	16%

Mathematical Strands of Proficiency

The two dominant strands of mathematical proficiency were conceptual understanding and procedural fluency. The category of *productive disposition* did not receive any codes, and therefore will not be discussed as part of the results section. The categories of strategic competence and adaptive reasoning both received codes from all sources proportional to the volume of codes produced by each. The relative frequency of these codes was much lower than the two dominant coding categories with 12% of the codes related to adaptive reasoning and only 3% related to strategic competence.

Textbook. Slightly more than half of all tasks and codes generated from these tasks originate directly from the textbook without modification by the instructor. The problems or tasks assigned after each lesson in the textbook followed a pattern related to the mathematical strands of proficiency. The initial exercises focused on a completion of one-step, basic skill, such as finding the probability associated with a z-score in a table.

These tasks were coded as *procedural fluency*. The next grouping of tasks focused solely on understanding a concept, such as the interpretation of the probability associated with a *z*-score, and involved no computation or procedures. These tasks were coded as *conceptual understanding*. The next group of tasks required both an understanding of the concept, choice and justification of an appropriate procedure, completion of the procedure, and an interpretation of the results. These tasks received multiple codes of *conceptual understanding*, *adaptive reasoning,* and *procedural fluency*. As described by the authors of the National Research Council (2001), *adaptive reasoning* is "the capacity for logical thought, reflection, explanation, and justification" (p. 116). Therefore, the choice accompanied with a justification of a procedure was coded as *adaptive reasoning*.

The relative frequency of these multidimensional tasks was low in comparison to other tasks, with only 15% of assigned textbook problems requiring *adaptive reasoning* or *strategic competence* in addition to *procedural fluency* or *conceptual understanding*. However, these types of tasks were present in most sets of exercises assigned by the instructor from the textbook and appeared to be transitional in nature as complex procedures were practiced for the first time by the cohort. The remainder of exercises typically consisted of practicing procedures and occasionally interpreting the results. These tasks were coded as procedural fluency with conceptual understanding if the task required the result to be interpreted beyond simply providing a numerical value.

Computer simulations. The computer simulation tasks comprise a relatively small percentage of the overall tasks and codes, 4% and 6% respectively. Since the computer simulation tasks consisted of running predetermined programs, these tasks were highly structured in nature. The computer simulations served a dual purpose. In order to

run the simulations, the cohort needed to have some familiarity with how to manipulate the software program to run the correct simulation or generate the required results. Once the simulation was complete, the cohort was asked to interpret the results. Simulation tasks were coded as procedural fluency for the steps required to run the simulation and conceptual in nature for the interpretation portion of the task. The results of coding the computer simulation tasks for mathematical strands of proficiency produced a balance between *procedural fluency,* 44% of codes assigned to this category, and *conceptual understanding,* 39% of codes assigned to this category.

Quantitative Literacy Series. The tasks completed from the *Quantitative Literacy Series* followed a similar pattern, but the emphasis on conceptual understanding was more pronounced. The tasks generally required the cohort to execute a series of procedures in order to generate results. The questions posed regarding the results required multiple interpretations of results, leading to heavy emphasis on conceptual understanding with 66% of the assigned codes in this category.

Instructor-generated. The tasks that originated from the instructor during lecture tended to be a balance of procedural fluency and conceptual understanding with a lesser amount of emphasis placed on adaptive reasoning and strategic competence. The percentage of codes assigned for each strand of mathematical proficiency are: (a) adaptive reasoning (8%), (b) strategic competence (1%), (c) conceptual understanding (39%), (d) procedural fluency (52%). In comparison to the course level results for all tasks, the instructor-generated tasks aligned closely with less emphasis placed on adaptive reasoning, 8% compared with 12%, and a higher emphasis placed on conceptual understanding, 39% compared with 33%.

Exams. In contrast to the instructor-generated tasks, course exams tended to focus heavily on procedural fluency, with 74% of the codes in this category compared to 52% at the course level. Less emphasis was placed on adaptive reasoning, 7% compared with 12%, strategic competence, 2% compared to 3%, and conceptual understanding, 18% compared to 32%, on exams than in the course overall. Therefore, the exams tended to emphasize procedural fluency over other mathematical strands of proficiency.

Chronological View. Shifting the view from the source of tasks to a chronological view of how the emphasis placed on each mathematical strand of proficiency changed overall throughout the course, Figure 4.32 conveys the frequency of codes with the four main topics of the statistics course.

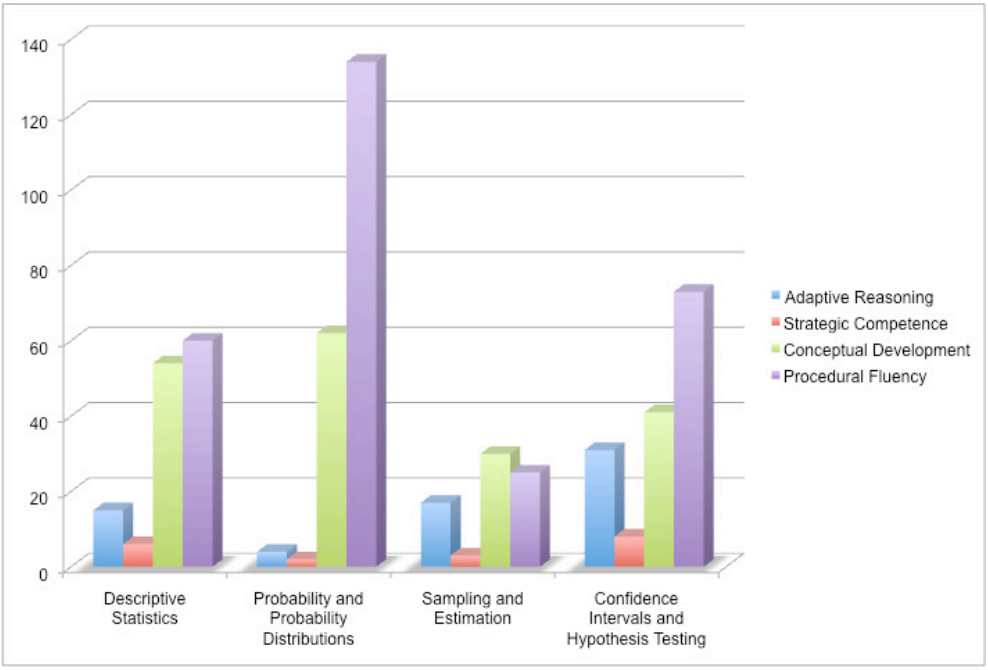


Figure 4.32. Fequency of coded tasks by mathematical proficiency strands

From this visual representation of how demands change throughout the course, several trends emerge. The most relevant trend related to inferential reasoning is the growing

requirement of adaptive reasoning throughout the course. However, the trend is relatively small in comparison to the other strands of mathematical proficiency with adaptive reasoning comprising only 12% of the overall codes. According to the Kilpatrick, Swafford and Findell (2001):

> Adaptive reasoning refers to the capacity to think logically about the relationships among concepts and situations and to justify and ultimately prove the correctness of a mathematical procedure or assertion. Adaptive reasoning also includes reasoning based on pattern, analogy or metaphor. (p. 170)

During the first half of the statistics course, on average one or two tasks per homework assignment include a component of adaptive reasoning out of approximately 20 assigned problems. However, in the later portion of the class, the number of tasks requiring this ability increases to 9 out of every 20 assigned problems and becomes critical to successful completion of inferential tasks in both homework assignments and assessments.

A lesser trend is demonstrated by the growth of the *strategic competence* strand throughout the course. However, the requirements in this regard are still fairly minimal, with 1 of 10 inference tasks requiring strategic competence in comparison to 1 of 23 earlier in the course. It is important to note that methods for solving statistical tasks were almost always explicitly prescribed. Strategic competence refers to the ability to formulate mathematical problems, represent them and ultimately solve them (Kilpatrick et al., 2001). Occasionally, the cohort is asked to choose between several available procedures. However, these decisions do not embody the intent of the strand of mathematical proficiency. Consequently, there were limited opportunities to acquire

strategic competence in the course. While informal approaches to generating an inference have the potential to align with the strategic competence strand, at no point in the course was the cohort asked to solve an inferential task with an informal approach. The tasks that were coded in this category related to generating valid samples, randomization, experimental design, one unstructured probability task, and modeling the inputs of formal hypothesis testing procedures in hypothetical situation.

The other categories of conceptual understanding and procedural fluency remain prominent throughout the course. A spike in procedural fluency is evident in the probability portion of the course and corresponds to tasks specific to combinations, permutations and multiplication rules.

In summary, the task analysis outlines the cohort's opportunity to learn inferential reasoning. The emphasis placed on the mathematical strands of proficiency throughout the course depicts the type of thinking advocated by the instructor and the supporting course materials. The data analysis identifies an imbalance of procedural fluency and conceptual understanding, thereby dwarfing the areas of strategic competence and adaptive reasoning. Therefore, the course primarily supported the cohort's comprehension of statistical concepts and relationships, as well as, their ability to accurately, efficiently and appropriately use procedures.

Content Analysis

The intent of the content analysis is to identify how the material was presented in terms of the order, the effort expended, and integration of topics. Figure 4.33 provides a summary of the content covered during each of the four main components of the course, in chronological order.

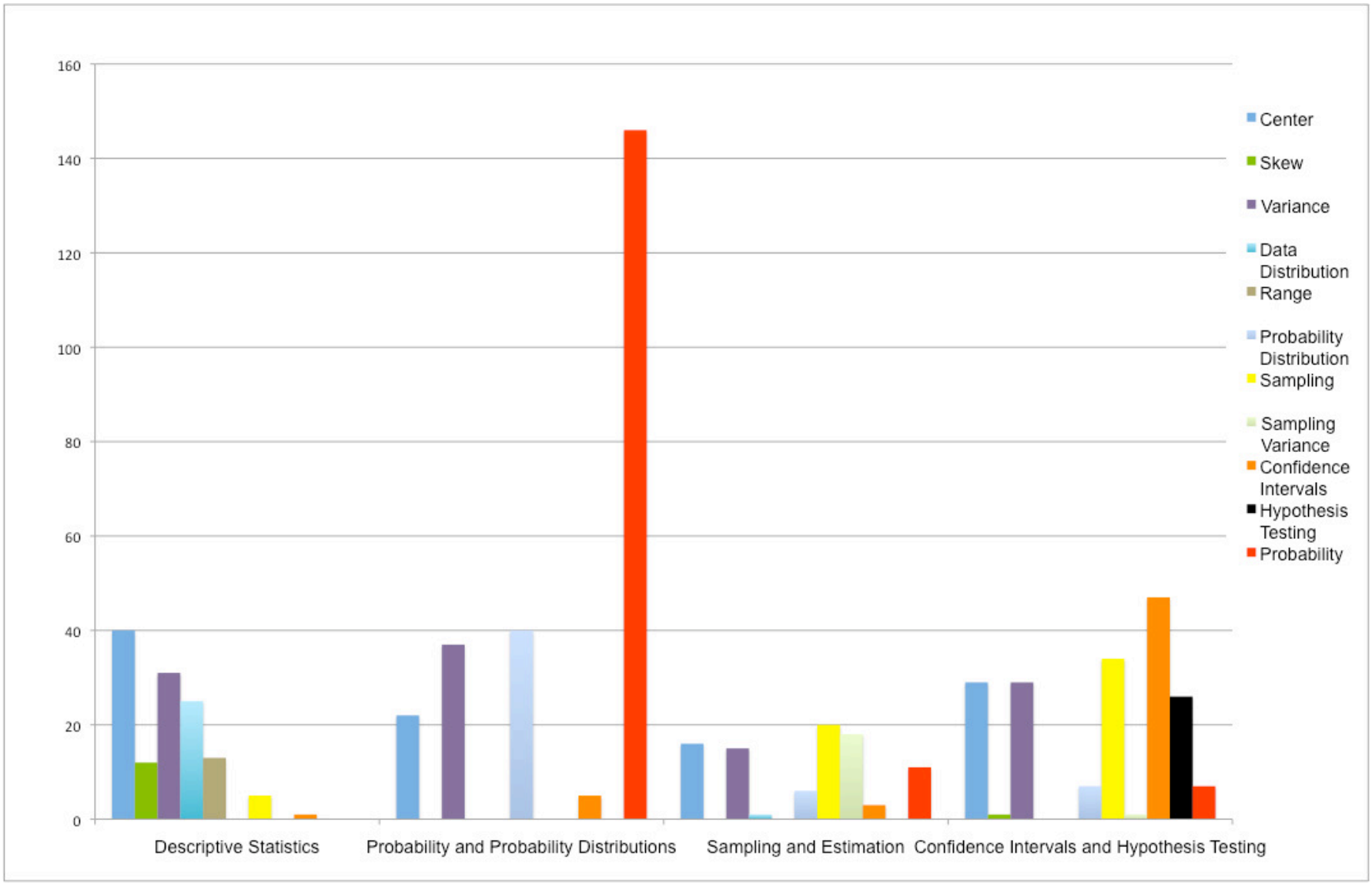


Figure 4.33. Frequency of coded tasks by content area

Consistent with the conceptual framework, the content areas represented are comprised of mainly of *core statistical concepts* and *aggregate statistical concepts* and the two formal methods of confidence interval and hypothesis testing. One additional category, probability, was added due to a large portion of the course tasks dedicated to this topic. The large number of probability tasks includes tasks such as finding the probability of simple events, combinations, and permutations. While a significant portion of the class, was spent on probability tasks, 38% of the overall content codes were assigned to probability, the conceptual framework for this study does not include this content as integral to the development of inferential reasoning. Figure 4.34 shows the same data with the probability tasks removed.

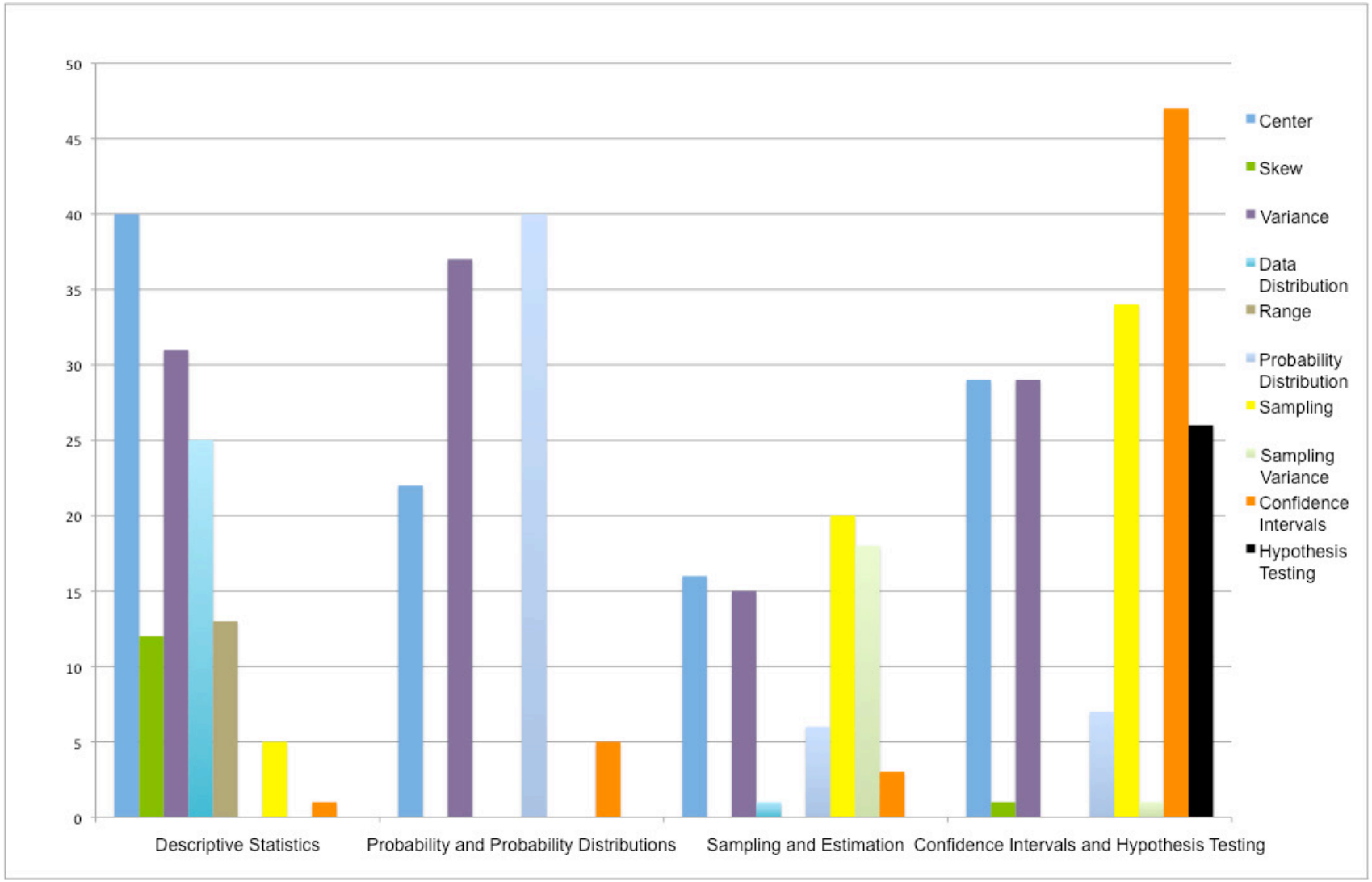


Figure 4.34. Frequency of coded tasks by content area without probability

As depicted in Figure 4.34, the course begins with tasks focused on the core concepts of measures of center, skewness, variability and range. In addition, the aggregate concept of data distribution also appears at the beginning of the course, which can serve as a unifying structure for building relationships between the core conceptions. The core concepts of skewness and range appear only in the first section of the course related to descriptive statistics and then quickly diminish. However, measures of center and variance remain prominent throughout the course and even increase in the last section related to the inferential methods of confidence interval and hypothesis testing.

The other aggregate concepts, sampling and sample variance, begin to appear in the third portion of the course related to sampling and estimation, comprising 42% of the content codes for the third portion of the course, as do the formal methods of confidence intervals and hypothesis testing, comprising 3% of the content codes. Sampling variance appears only in the third section, while sampling grows in absolute terms from appearing in 30 tasks during sampling and estimation to 42 tasks in the fourth portion of the course related to inference.

During the fourth and final portion of the course, the inferential method of confidence interval grows substantially and comprises 26% of the content codes for this portion of the course. The number of tasks associated with confidence intervals is nearly double the number of tasks focused on formal hypothesis testing, as hypothesis content codes only comprise 14% of the total codes during this course component. Confidence interval methods appear earlier in the textbook than hypothesis testing. While the associated chapter in the book was completed for confidence intervals, the hypothesis test chapter was not completed in its entirety by the end of the course. In addition, 19 in-class

tasks from the *Quantitative Literacy Series* were dedicated to generating confidence interval inferences, compared to 0 tasks for hypothesis testing.

Core concepts. The emphasis placed on the core concepts throughout the course is shown in Figure 4.35. The vertical axis provides the total number of content code assigned for each core concept for the tasks contained within the statistics during the main sections of the course.

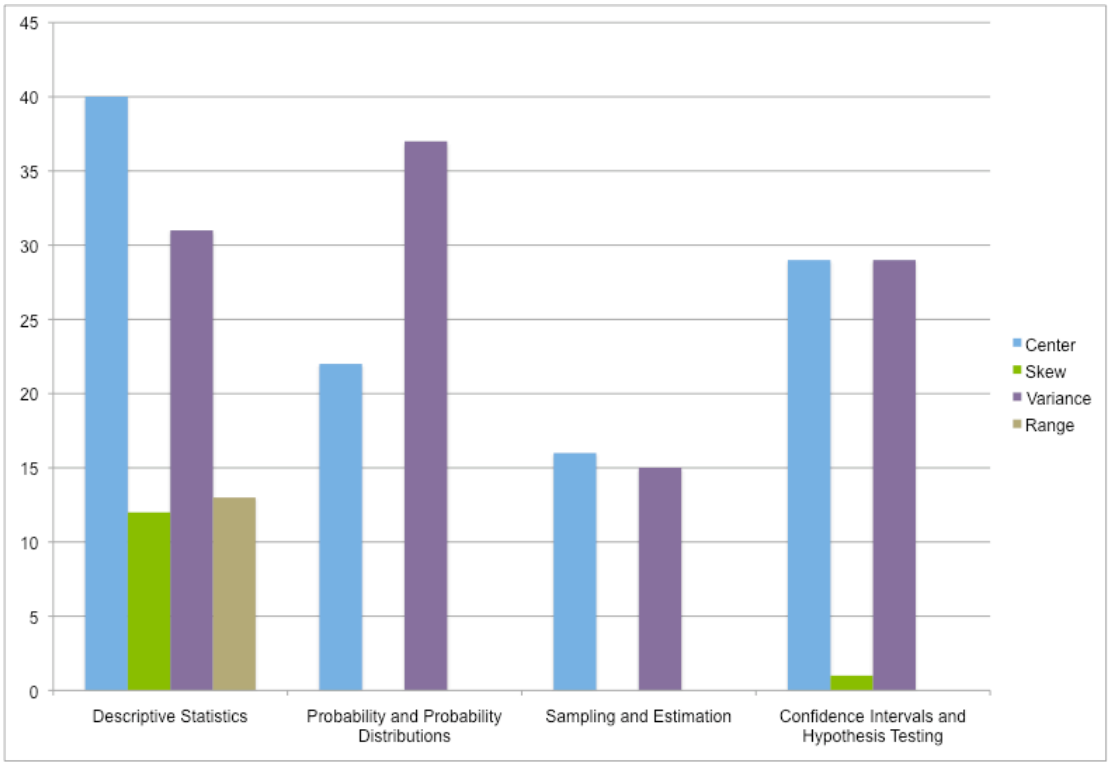


Figure 4.35. Frequency of coded tasks by core concepts

The first portion of the class, descriptive statistics, includes graphing. The cohort is asked to create visual displays of data such as bar charts, stem and leaf graphs, and dot plots. The data distributions are described in terms of measures of center, skew and range. Once graphing is completed in the textbook in the first chapter, graphs of data rarely appear again in the textbook. In total, 10 tasks involve either the creation of graphs or interpretation of graphs in the first chapter, but no other graphs are used in tasks

afterwards. Box plots are addressed in a later chapter, but box plots have the potential to obscure important details of the shape of the data distribution. Data are described mainly in terms of the mean or proportion, standard deviation, and relationship to probability distributions. Hence, visual representations of core concepts such as skew or range are not present in tasks as graphical displays of data are not provided. The consistent focus on measures of center and variance is explained by the fact that data distributions are described in these two terms throughout the course.

Aggregate concepts. The emphasis placed on the aggregate concepts and formal inferential methods throughout the course is shown in Figure 4.36.

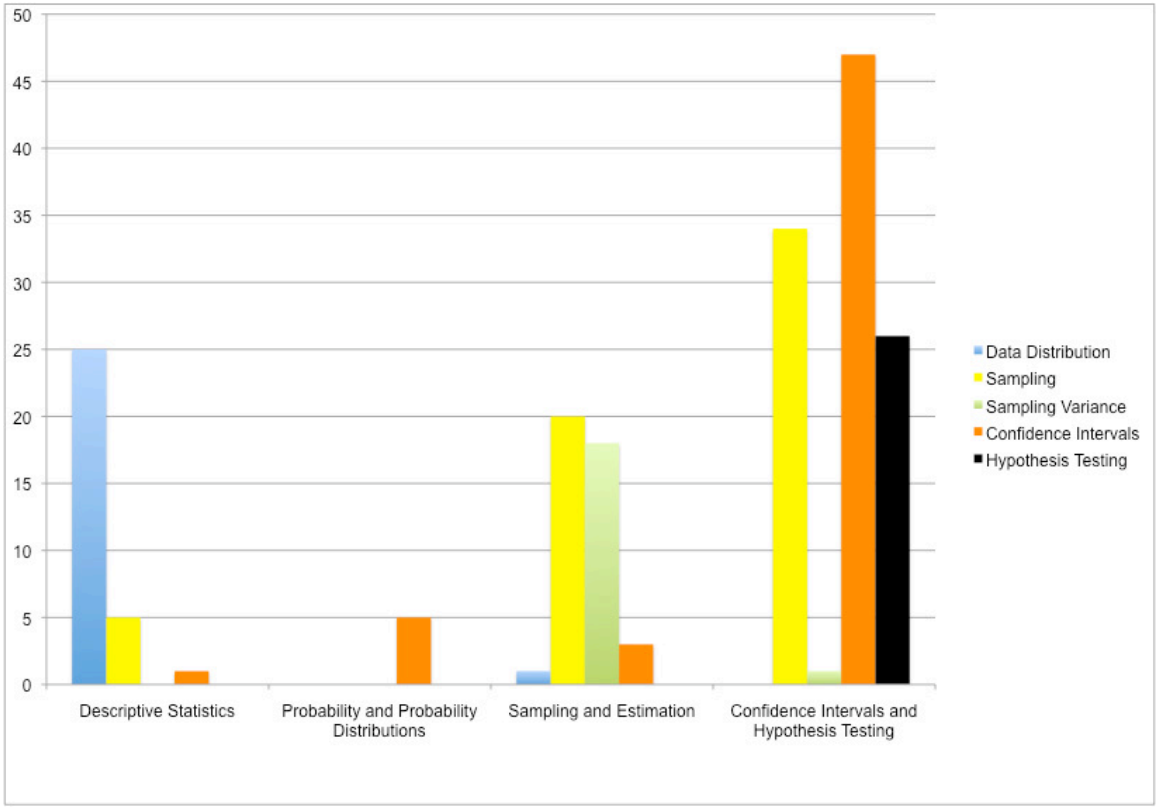


Figure 4.36. Frequency of coded tasks by aggregate concepts

Since the textbook chapter for confidence intervals was completed whereas the chapter for hypothesis testing was not, more course tasks were dedicated to confidence

interval approaches than formal hypothesis methods. Sampling tasks comprise a larger portion of the content in the last half of the course and contain problems related to obtaining an adequately large sample, designing a survey to obtain an unbiased sample, estimating population characteristics from samples, determining the relationship between samples and sample distributions, and creating inferences about a larger population from samples. The transition from sampling to determining a confidence interval for random variables with normal, Poisson and binomial distributions are attended to in chapter 8. Chapter 9 was partially completed and two hypothesis tests were introduced and discussed: the difference in means test, and the difference in proportions test.

In summary, the opportunity to learn as characterized by a content analysis of the tasks demonstrates that core conceptions of center and variability and the aggregate conception of sample were well represented in the tasks for the statistics course. The core conceptions of skewness and range appear primarily in the first portion of the course, as does the aggregate conception of data distribution. Sampling variability is relatively unattended to throughout the course. The formal method of confidence interval is developed throughout the latter portion of the course and the number of tasks requiring this method is twice as large compared to hypothesis testing tasks. Hypothesis testing is covered near the end of the course. Incorporating the findings from the classification of tasks by mathematical strands of proficiency, an increased demand for adaptive reasoning occurs simultaneous with the introduction of formal inferential methods. Prior to this increase in adaptive reasoning demand, the primary strands are conceptual understanding and procedural fluency. Therefore, the content associated with measures of center, variability, data distribution and sampling was taught with a focus on concepts and

procedures. However, the inferential methods were approached with the additional element of adaptive reasoning, although the focus on conceptual understand and procedural fluency remained dominant in relative terms.

Results Summary

The findings of this study reveal that over the duration of the statistics course a slight majority of the cohort improved their ability to informally inferentially reason, generally by one level as defined by the SOLO taxonomy. A smaller percentage remained at the same level. Only a few preservice teachers exhibited lower levels of inferential reasoning at the end of the course, utilizing primarily a frequency approach instead of employing sound proportional reasoning. Those whose dominant level of reasoning changed from Unistructural to Multistructural tended to incorporate either variance or spread into their responses in addition to measures of center. Those who remained at the Unistructural level did not incorporate concepts beyond measures of center into their reasoning, although interviews revealed that some participants at the Unistructural were in fact noticing variance but elected not to include this element in their responses. Preservice teachers who performed at a Prestructural dominant level of inferential reasoning exhibited difficulty in proportional reasoning and regularly offered opinion statements rather than data-based arguments. Finally, the preservice teachers who reasoned at the Relational level incorporated all relevant aspects of tasks both in terms of statistical information and context to provide a comprehensive and logical response. In addition, those who provided formal approaches to generate inferences on the post-assessment tended to maintain their level of reasoning regardless formal or

informal approaches. When formal results conflicted with informal reasoning, the cohort tended to default to informal reasoning.

Analysis of preservice teachers' opportunity to learn indicates that while some core and aggregate statistics concepts such as mean, variance and sampling received regular attention throughout the course, other concepts received sporadic attention including data distribution, skew, range and sampling variability. In addition, the formal method of using confidence intervals to generate inferences was covered more thoroughly than formal hypothesis testing. Probability content comprised a significant portion of the course but was largely unrelated to generating inferences or supporting the underpinnings of inferential reasoning. The mathematical strands of proficiency analysis revealed little expectation to develop adaptive reasoning, at least until the end of the course, and even less demand was placed on the development of strategic competence. On the other hand, throughout the course, a consistent level of conceptual development and procedural fluency was maintained.

In the next chapter, I provide a discussion of results and offer implications for middle and secondary school teacher preparation programs and curriculum developers. In addition, I discuss the limitations of this study and suggest an agenda for further research topics that may result from this endeavor.

CHAPTER 5: DISCUSSION, SUMMARY, AND RECOMMENDATIONS

Over the past century, statistics has grown from "relative obscurity in the mathematics curriculum to important, fundament topics that should be studied by all students at each grade level" (Jones & Tarr, 2010, p. 73). With regard to statistical inference, current national standards for middle and secondary statistics education advocate for the introduction of inferential reasoning during middle school through informal approaches, followed by formal approaches in secondary years (Franklin et al., 2007; NGA and CCSSO, 2010). Research indicates that little is known about the extent of knowledge needed to effectively teach statistics in middle and secondary school settings (Garfield & Ben-Zvi, 2008), in part because few investigations have been conducted with preservice teachers who have had an opportunity to learn statistics content (Shaughnessy, 2008). Accordingly, this study seeks to address this void in the literature by characterizing a cohort of middle and secondary mathematics preservice teachers' inferential reasoning while enrolled in a statistics course designed specifically for future teachers. In addition, a relationship is sought between the cohort's change in informal and formal inferential reasoning. In order to provide a context for such changes, the cohort's opportunity to learn inferential reasoning is characterized through a careful analysis of all tasks contained within the course.

In this study, I address the following research questions:

1. How can the change in middle and secondary preservice teachers' inferential reasoning abilities be characterized during a statistics course?

2. Does a relationship between preservice teachers' change in informal and formal inferential reasoning exist? If so, how can it be characterized?
3. What opportunities to learn inferential reasoning are afforded middle and secondary preservice teachers during a semester-long statistics course?

In this chapter, I summarize the study and discuss the findings. The chapter is organized into five sections: (a) a summary of the study and findings, (b) a discussion of findings, (c) limitations of the study, (d) implications for teacher education, and (e) recommendations for future research.

Summary of the Study and Findings

Researchers have found that despite advancing statistics content recommendations, actual changes in classrooms have lagged behind (Jones et al., 2007). One possible reason for this is that teachers have not themselves experienced learning statistical content in alignment with current standards. However, knowledge of the statistical content to be taught is clearly an essential component. Until recently, many middle and secondary school mathematics teachers have not had an opportunity to learn statistics content during their college coursework (Shaughenssy, 2007). This study provides a needed assessment of how middle and secondary mathematics preservice teachers inferentially reason *before* and *after* a statistics content course.

Methodology

The participants (n=33) for this study are middle and secondary mathematics preservice teachers enrolled in a statistics course designed for future teachers at a large Midwestern university. In order to characterize the preservice teachers' change in inferential reasoning, three assessments were administered at key points in the course.

The entire cohort completed a pre-assessment during the first day of the course; representing a wide range in statistical thinking, a stratified random sample (n=12) participated in midcourse clinical interviews; and the entire cohort completed a post-assessment during the final week of the course. The assessments consisted of tasks designed to elicit inferential reasoning responses based on prior research studies (e.g., Bakker, 2004; Cobb, 1999; Garfield et al., 2007; Watson, 2002) and embodied the following critical attributes of assessment tasks: ill-structured, open-ended, represented visually, and embedded within a relevant context.

Responses to assessment tasks were classified in accordance with a hierarchical, cognitive framework rooted in neo-Piagetian, cognitive psychology and especially adapted for statistical reasoning (Biggs & Collis, 1982). Informal inferential reasoning approaches represent the concrete mode of reasoning, and formal approaches represent the formal mode. Informal responses to tasks were assigned to one of four reasoning levels using a modified SOLO taxonomy informed by prior research (Mooney, 2002), and formal reasoning responses were also assigned to one of four reasoning with the original SOLO taxonomy. For each assessment, responses to tasks were converted to an interval scale so that means and dominant levels of inferential reasoning could be determined. When the mean fell between two reasoning levels, the frequency of response types were examined in order to determine the modal reasoning level for a given preservice teacher. The change in the preservice teachers' inferential reasoning was characterized by reporting differences between dominant levels of inferential reasoning from the pre- to post-assessments at the individual level.

In addition to characterizing the overall change in the cohort's inferential reasoning, a comparison of the cohort's informal and formal inferential reasoning was undertaken to determine if a relationship existed between the two reasoning approaches on the post-assessment. A dominant *informal* inferential reasoning mode was determined for each preservice teacher using the same procedure described previously. Because the cohort responded to only a few tasks with formal responses, this *informal* dominant mode of reasoning was compared to each *formal* response provided to determine whether response pairs were *concordant* or *discordant*. In this manner, a correlational measure was generated that represents the relationship between the cohort's informal and formal inferential reasoning on the post-assessment.

Lastly, the cohort's opportunity to learn inferential reasoning during the statistics course was characterized through an analysis of all tasks (n=375) utilized in the course including homework problems assigned from the textbook, tasks generated by the instructor during lecture, computer simulation activities, in-class group projects, and course assessments. The tasks were analyzed in relation to two important dimensions: statistical content necessary to complete the task, and the *strands of mathematical proficiency* (Kilpatrick et al., 2001).

Results of the Study

Preservice middle and secondary mathematics teachers' inferential reasoning. Responses to the three assessments provide evidence of growth in inferential reasoning at the class level from the beginning to the end of the statistics course. From the pre- to post-assessment, the average dominant level of reasoning for the cohort as a whole shifted from Unistructural in nature, attending to one correct data comparison in order to

generate an inference, to Multistructural in nature, incorporating several relevant aspects. However, considerable variation at the cohort level *within* specific tasks, and at the preservice teacher level *across* tasks was evident on both assessments.

At the *cohort level*, the variation in reasoning specific to tasks is measured by the standard deviation in levels of inferential reasoning, and ranged from 0.35 to 0.99 on the pre-assessment and 0.79 to 1.08 on the post-assessment. Tasks with standard deviation values below 0.75 tended to elicit only one or two different levels of inferential reasoning compared to tasks with standard deviation values near 1.0 that tended to elicit the full spectrum of possible levels of inferential reasoning. One task that appeared on both the pre- and post-assessment was ultimately excluded from data analysis because its context may have yielded responses that were not representative.

At the *preservice teacher level*, stability in inferential reasoning across tasks was analyzed in order to determine a dominant level of inferential reasoning for each participant on each assessment. Preservice teachers with the highest degree of variation across tasks tended to exhibit responses at three or more levels of inferential reasoning. In contrast, the preservice teachers with the lowest level of variation provided inferential reasoning responses at only two levels of reasoning.

Prior to the course, the cohort reasoned predominantly at the Unistructural level because responses typically consisted solely of informal approaches. More specifically, the portion of the cohort at the Unistructural inferential reasoning level generated inferences primarily based on changes in measures of center. Those at the Prestructural level either did not provide data-based arguments or neglected to apply proportional reasoning. The relatively small number of preservice teachers at the Multistructural

inferential reasoning level attended to both changes in measures of center and one other aspect of the data. At the end of the course, the cohort's inferential reasoning was spread across the four possible levels, with most at the Multistructural level and demonstrating both informal and formal approaches to assessment tasks. The preservice teachers at the Relational inferential reasoning level provided qualitatively different responses that fully considered all data provided in the tasks and the context of the task in order to generate an inference.

The changes in dominant levels of inferential reasoning differed between the middle school and secondary preservice teachers. Although 58% of all participants increased their reasoning ability, the growth was more pronounced for secondary teachers with 75% increasing one or more levels of inferential reasoning compared with 50% for the middle school population. In addition, 12% of the cohort experienced a one-level decline in inferential reasoning, most of whom were middle school preservice teachers. The amount of preservice teachers who remained at the same level of inferential reasoning was similar between the two groups with 30% overall, 31% specific to the middle school population and 29% for the secondary. While the middle school preservice teacher population reported completing more prior statistics coursework than their secondary peers, comparable gains in inferential reasoning were not realized. In addition, the secondary preservice teachers completed more advanced mathematics coursework than the middle school preservice teachers and more mathematics courses overall.

Commonalities in the inferential reasoning responses were discerned for groups of preservice teachers, who reasoned at the same level, advanced one or two levels and

decreased by one level. Of those who remained at the same dominant level of reasoning, the largest portion remained at the Unistructural level. These preservice teachers primarily focused on measures of center when generating inferences and did not coordinate information related to other core statistical concepts such as range of the data distribution and variation in results. Midcourse interview data suggest that some preservice teachers at this level of inferential reasoning noticed other elements in the data distribution but did not incorporate this information into responses. Several secondary preserivce teachers remained at the Multistructural level of inferential reasoning. Typical responses provided by these preservice teachers accounted for differences in measures of center and attended to one additional global characteristic of the data distributions such as spread or variation. The responses tended to fall short of Relational levels of inference because they did not consider all data provided or presented justfications in a disjointed and sequential fashion rather than an integrated whole.

The most common change was a one-level increase with the portion advancing from the Unistructural level of inferential reasoning to Multistructural. Typical responses at both levels of reasoning follow the examples provided previously. Two middle school preservice teachers changed from Prestructural to Unistructural as their dominant level of inferential reasoning. However, opinion statements, additive reasoning and attention to local events such as outliers continued to appear sporadically in responses. The five preservice teachers, whose dominant level of inferential reasoning changed from Multistructural to Relational, represented the highest performing group. Characteristic trends in responses included attending to all data provided in tasks, coordinating core and aggregate statistical concepts to arrive at a comprehensive inference, and providing

interpretations anchored in the task context. In addition, the Relational responses quantified changes between data distributions in order to justify significance or the lack of signficance. Lastly, two preservice middle school teachers declined one level in inferential reasoning moving from the level of Unistructural to Prestructural. These two preservice teachers shared characteristics with those who began at the Prestructural level.

Specific to formal inferential reasoning, results followed the trends exhibited in the post-assessment overall. The cohort tended to reason between the Unistructural and Multistructural levels, but closer to Multistructural. Several common errors observed on formal responses include: interpreting the statistical meaning of significance, selecting an appropriate formal method, and populating formula values incorrectly.

Relationship between informal and formal inferential reasoning. Results from the post-assessment support the existence of a relationship between informal and formal approaches to inferential tasks. Given that the frequency of formal responses to tasks was low in comparison to informal responses, each formal task response level was compared to the associated individual preservice teachers' dominant levels of informal reasoning on the post-assessment. For a given preservice teacher, 80% of levels assigned to formal inferential task responses were concordant with the dominant informal inferential reasoning level. Therefore, a substantial relationship appears to exist as preservice teachers tended to reason at similar levels for both formal and informal approaches on the post-assessment.

The opportunity to learn inferential reasoning. The cohort's opportunity to learn inferential reasoning is characterized through an analysis of the 375 tasks contained within the statistics course. The content analysis was conducted in accordance with the

conceptual framework and determined that the core conceptions of center and variability and the aggregate conception of sample were well represented in the statistics course. The core conceptions of skew and range appear primarily in the first portion of the course, as does the aggregate conception of data distribution. However, the concept of sampling variability is sparsely represented throughout the course. The formal method of confidence interval is developed throughout the latter portion of the course, and the number of tasks requiring this method is twice as large compared to hypothesis testing tasks. Hypothesis testing is not taught until the end of the course, in the final chapter of the textbook. Graphical displays of data appear in the first chapter of the course related to descriptive statistics, but are largely not present afterwards.

Incorporating the findings from the classification of tasks by mathematical strands of proficiency (Kilpatrick et al., 2001), an increased demand for adaptive reasoning occurs simultaneously with the introduction of formal inferential methods with 37% of inferential tasks requiring adaptive reasoning versus 12% for other tasks in the course. Prior to the topic of inferential reasoning, the primary proficiency strands emphasized by tasks are conceptual understanding (56%) and procedural fluency (75%). Therefore, the content associated with measures of center, variability, data distributions and sampling was taught with a focus on concepts and procedures. However, the inferential methods were approached with the additional element of adaptive reasoning, although the focus on conceptual understanding and procedural fluency remained dominant in relative terms.

Discussion of Findings

The purpose of this study was to characterize changes in the preservice teacher's inferential reasoning while enrolled in a statistics course. However, the identification of specific catalysts for learning and changes in inferential reasoning was beyond the scope of the study.

The most prevalent trend was a one-level change in inferential reasoning, from Unistructural to Multistructural on the pre- to post-assessment, respectively. Responses provided at this level tended to be based on changes in measures of center or shifts in modal clumps and add a global comparison between data sets. Watson (2003) notes in a study across grade levels 3, 5, 6, 7, and 9 that reasoning with variation was not a good predictor of the successful inferential reasoning, but rather that noticing the shape of the data displays assists more strongly in effective decision-making. In this study, responses to tasks incorporated a variety of global characteristics beyond measures of center in order to advance inferential reasoning, albeit in a disjointed and uncoordinated manner. Hence, findings support that noticing the shape of data displays assists more strongly in decision-making, than attending specifically to variation. Examples of global comparisons include range, spread and variance, and these topics were indeed given substantial attention in the first half of the statistics course.

Of interest is the cause for these preservice teachers to attend to additional statistical concepts from the beginning to end of the statistics course. Watson (2003) found that student's inferential reasoning could be advanced through the introduction of cognitive conflict in a similar manner to that achieved through maturation. Given that participants in this study were of adult age, one would expect that interactions in the

environment that supported the learning of inferential reasoning. The statistical content referenced in Multistructural responses tended to be addressed in the first portion of the course. The concepts of center, range, variance, distribution and skew were discussed in first chapter of the statistics course through the use of data displays. Therefore, it is plausible that the coupling of visual data displays with an emphasis on concepts related to data distributions provided the necessary opportunity to learn for a portion of the cohort who attended to more than just measures of center.

Another possible explanation for the growth in reasoning may relate to the explicit pairing of *mean* with *variance* in course tasks. Results of the content analysis revealed that during the last half of the course, rarely was the mean for a data set provided without the variance or standard deviation. Therefore, through these tasks a consistent message was implicitly communicated that *both* concepts were needed to describe the data set and generate an inference, and this may have stimulated awareness that *two* pieces of information are needed in justifications.

A remarkable finding of this study is that no preservice teachers demonstrated Relational inferential reasoning prior to the statistics course. This is somewhat surprising given that two-thirds of the cohort had completed either an introductory or advanced statistics course prior. This finding suggests that carefully structured learning interactions are required for individuals to advance reasoning to the Relational level. In other words, maturation and unstructured life experiences may not promote full development of inferential reasoning ability. Since inferential reasoning is fraught with a number of misconceptions relating to sampling (Tversky & Kahneman, 1971), the need for authentic experiences in sampling data and developing a sense of expected variance in

samples is critical for the process of generating inferences. In the content analysis, students rarely engaged in authentic sampling of data, and thus misconceptions may not have been elicited nor confronted. In addition, the course tasks did not require adaptive reasoning to a substantial level prior to the latter portion of the course. These two omissions might explain why so few preservice teachers attained the Relational level of inferential reasoning.

Only a few preservice teachers, three secondary and two middle school, ascended to the Relational level by the end of the course, which is consistent with prior research findings that report disappointing results from college level statistics courses related to interpreting results of hypothesis tests (e.g. delMas et al., 2007). While the pre-assessment dominant reasoning levels of these five preservice teachers were slightly higher than the cohort's average, the only notable difference in the group is that one member had previously completed an AP statistics course. Why these five were able to reason at the Relational level and others in the cohort were not by the end of the statistics course is worth further exploration.

The five preservice teachers at the Relational level demonstrated flexibility in terms of utilizing both formal and informal approaches on the post-assessment. Because preservice teachers who reasoned at the Relational level responded with both informal and formal responses to inferential tasks, this study provides a needed distinction between those who can *carry out* formal procedures and those who *understand* the process of statistical inference. Previous research indicates that the ability to successfully complete a formal hypothesis test or formal process does not ensure an understanding of how to apply inferential processes to ill-defined, real-world context (Aquilonius, 2005;

Liu, 2005). In this study, many preservice teachers at the Multistructural level were able to execute formal methods effectively, by choosing appropriate tests and entering correct values for variables, but then fell short of interpreting the results by either claiming significance or lack of significance inappropriately. In addition, those who responded at the dominant Multistructural level of reasoning fell short of generating a well-supported informal inferential response, by failing to provide quantitative support for claims of significance. However, preservice teachers at the Relational level were able to correctly interpret results from hypothesis tests and also provide likewise sophisticated responses using informal approaches.

In general, supporting or refuting significance posed a challenge for preservice teachers below the dominant level of Relational inferential reasoning. As an example, on the Speed Trap task many preservice teachers stated that a 1.8 miles/hour reduction in the average speed of 60 cars was insignificant without providing any additional justification. Prior research has found that upon completion of an introductory, college level, statistics course, the least understood concept is significance, and that most students only discuss significance through relationships to other topics (Williams, 1999). For those at the Prestructural and Unistructural dominant reasoning levels, judgments regarding significance tended to either consist of mere opinion statements or were based on claims about absolute (not relative) differences in means without consideration for variance or sample size. For these preservice teachers, significance was determined by the context of a given task and absolute differences of only a single measure. One possible explanation for the difference between those who pursued quantitative approaches to significance versus those who did not is that the former group may understand the reasoning behind

the steps necessary to generate an inference (collecting a sample or samples, accounting for possible error, determining significance based on p-values), while the latter group does not. The heavy emphasis placed on procedural fluency versus adaptive reasoning in the course tasks strongly suggests that carrying out algorithms was privileged over understanding why procedures work and how variables relate to generate the an inference.

A key finding of this study is that a considerable portion of the cohort did not advance their inferential reasoning from the beginning to end of the course. The need for the discipline of statistics arises from the omnipresence of variability (Moore, 1997) and the associated need to attend to variation (Wild & Pfannkuch, 1999). In contrast to the findings of Konold and Pollatsek's (2002) study, preservice teachers who reasoned at the Unistructural level typically attended to measures of center, not variation. In fact, a large portion of the cohort reasoned predominantly with measures of center, as 70% reasoned at the Unistructural level on the pre-assessment and 33% on the post-assessment. Archetypal responses to inferential tasks on both assessments focused on changes in measure of center in the form of either absolute differences in means or estimated shifts in modal clumps of data. Such results are consistent with research conducted with inservice secondary mathematics teachers, which found that teachers are generally comfortable comparing distributions informally with descriptive statistical measures such as mean, but struggle to reason about multiple types of variation both *within* and *between* data distributions (Makar & Confrey, 2001). As previously stated, those who reasoned at the Unistructural level, as evidenced by the case of Dave, may have noticed variation within and between data distributions but nevertheless neglected to include these in

written responses. Since Dave and others were not queried about deficiencies in their reasoning, the reasons behind such omissions are unknown. One possible explanation is that those who reasoned at the Unistructural level either believed that these observations were not relevant or they were not desired, given their prior experiences in deterministic approaches to statistics that emphasize measures of center (Leavy, 2010). Another possible factor may have been the lack of prior advanced mathematics coursework taken by the middle school preservice teachers. The secondary school preservice teachers tended to reason at higher dominant levels by the end of the course in comparison to their middle school peers and most had completed three semesters of Calculus. The reasoning requirements of advanced mathematics courses may have provided experiences that prepared the secondary preservice teachers for justification and argumentation demands.

A common feature of all preservice teachers was the variation in their responses. On both the pre- and post-assessments, approximately one third of the preservice teachers exhibited three different levels of inferential reasoning on tasks, and this finding is consistent with the notion that students' reasoning is often inconsistent from test-item to test-item (Garfield & Ben-Zvi, 2007). The tasks posed on assessment were fairly similar in nature in terms of the statistical content but differed in their contexts and data representations. Therefore, the variation in inferential reasoning seemed to relate more to the preservice teachers' interpretation of the contextual considerations of the problem statements and data representations and values provided rather than the tasks themselves. Another one third of the cohort was unaffected by these differing data representations, values of data, and contexts, but did not perform at higher or lower levels of inferential reasoning. The reasons for the stability or instability in reasoning across items are

unclear, but this phenomenon was evident on both assessments. Moreover, those preservice teachers who exhibited the highest level of variation on the pre-assessment did not demonstrate as much inconsistency on the post-assessment, suggesting at the individual level the presence of considerable variation both *within* and *across* assessments.

Limitations of the Study

Limitations are inherent in all research studies, and three limitations of this research study are discussed below.

Nature of the Assessment Tasks

In order to fully assess preservice teachers' ability to reason inferentially, authentic tasks are needed which allow participants to define the problem statement, design an experiment, collect the necessary data and make appropriate interpretations. Although the tasks included in the assessments embody recommended attributes of high-quality tasks (e.g., ill-structured, open-ended), all data within the tasks were provided; stated differently, preservice teachers neither generated nor collected any empirical data. Hence, data collection choices and the associated implications were not assessed through the tasks to a large degree. In addition, the extent to which the preservice teachers were required to define the problem statements was limited in nature, as preservice teachers were asked to interpret and parameterize problem statements from an overarching goal. Wild and Pfannkuch (1999) advocate the need to expose students to statistical problems, which require them to make decisions regarding problem formulation and experimental design, as do the authors of the *GAISE* recommendations (Franklin et al., 2007). By

adding such elements to tasks, additional complexity would have been introduced and thereby making it more challenging to hone in on inferential reasoning.

In addition, the data sets included in assessments were almost exclusively represented as dot plots with only a few cases of box-plots and histograms. Therefore, the reasoning processes associated with generating inferential statements may under represent approaches associated with the comparison of data displayed in box-plots and histograms.

Formal Inference Characterization

With regard to formal inference, this study represents a first attempt to apply the SOLO taxonomy to responses of formal inferential tasks. The characterization of formal approaches to inferential tasks is based on three components: choice of the appropriate hypothesis test or formal method, execution of the methods with appropriate values for variables, and a reasonable interpretation of results. By focusing on these three components, the *reasoning* portion of formal methods was relegated to the final step of the process, namely the interpretation of results. By attempting to segregate informal and formal approaches, an element of *adaptive reasoning* was not assessed. The implications regarding the variables that serve as inputs to the formal method (mean, proportion, variance and sample size) and the relationships between these inputs were not evaluated.

Context of the Study

The results of this study may not be generalizable to all contexts, as they represent a specialized research context of one statistics class taught by a particular instructor. The class was taught once a week for a three-hour block, likely atypical of most statistics courses in the country, and used a specific set of course materials. The impact on

inferential reasoning of a different instructor, using alternative curricular materials, and offered in another course format is simply unknown.

Implications for the Statistical Preparation of Teachers

The results of this study provide insight into the preparation of middle and secondary mathematics teachers specific to statistical inference. In this section, I present implications for teacher preparation drawn from these results in the areas of adaptive reasoning, proportional reasoning, the role of probability, and informal inferential reasoning.

Greater Emphasis on Adaptive Reasoning

Through the task analysis, an imbalance in emphasis placed on the mathematical strands of proficiency exists in the statistics course with the strands of conceptual understanding and procedural fluency dominating the course experience. Adaptive reasoning—"the capacity for logical thought, reflection, explanation, and justification" (Kilpatrick et al., 2001, p. 116)—was emphasized only modestly and largely confined to the final components of the course. Because of the inconsistent focus on adaptive reasoning, the cohort may have been ill prepared to provide justification for choices made in generating inferences that coordinating multiple concepts as demonstrated on post-assessment tasks. Therefore, adaptive reasoning in the form of answering questions that begin with the word "why" is critical to developing argumentation schemas that organize conceptual and factual information. These schemas will later be called upon in order to generate an inference and organize new pieces of information. Means and Voss (1996) advocate that adaptive reasoning and argumentation must be developed over time and across disciplines. Therefore leaving this demand unaddressed until the end of the

statistics course creates a quantum leap of expectation for preservice teachers who were not accustomed to organizing their learning in this fashion.

Explicit Focus on Proportional Reasoning

Based on analysis of pre- and post-assessment responses, five middle school preservice teachers demonstrated a lack of proportional reasoning and instead favored absolute frequency approaches. Not surprisingly, these preservice teachers reasoned at the dominant levels of Prestructural or Unistructural throughout the course. In a cognitive conflict research study, Watson (2002) found that it is very difficult to change the inferential reasoning of those who favor absolute frequency approaches over proportional approaches. Van Dooren et al. (2005) identified structural components of mathematics tasks that cue secondary students to engage in proportional reasoning, and the tasks included within this study did not include such cues. Perhaps the structure of the tasks coupled with a rigid conception of *when to apply* proportion reasoning hindered a subset of the cohort. Due to the lack of positive change in inferential reasoning with these five middle school preservice teachers, results of this study support the notion that effective use of proportional reasoning is necessary for advancing inferential reasoning ability. Therefore, instruction should seek to assess whether students are appropriately reasoning proportionally early in statistics coursework. In addition, remediation is needed for those who demonstrate a preference for absolute frequency approaches for comparing sets of data or inconsistent use of proportional reasoning. The lack of proportional reasoning among preservice middle school mathematics teachers threatens the prospect of developing proportional reasoning in middle school students.

The Role of Probability: A Tool for Statistics

Probability is an essential tool for statistics and supports the process of generating inferences. However, probability tasks that focus on counting techniques such as combinations and permutations appear unrelated to generating inferences. If the goal of an introductory statistic courses is to teach inference, then all topics within the course should relate to that primary objective. An entire chapter of the course textbook, equating to two weeks of course time, was dedicated to applying combinations and permutations. In statistics courses, the time dedicated to content ultimately unrelated to generating inferences should be lessened in order to allow for key missing components, such as informal approaches to inference. In addition, probability tasks should not be taught in a stand-alone matter but instead should connect to other topics within the course to illustrate the role of probability as a tool for statistics.

Explicit Attention to Informal Inferential Reasoning

In order to prepare middle and secondary school mathematics preservice teachers adequately to teach statistics in alignment with current recommendations, preservice teachers need an opportunity to learn the same content that they will ultimately be teaching. Given the heavy emphasis of informal inferential reasoning in middle school, tasks such as these must be included in courses that are specifically designed for teachers. At no time during this course, was the cohort asked to use informal approaches to compare two data distributions and generate an inference. Therefore, informal inferential tasks need to be added to the curriculum of middle school mathematics preservice teachers, as this course represents preservice teachers' last opportunity to learn informal approaches to inference.

In addition to not aligning with current content recommendations, formal approaches were clearly privileged over other, less formal methods. Informal inferential methods should prevail early in the course and provide a bridge to introducing formal methods later parallel to the current mathematics curriculum standards, which place an emphasis on informal inference in the middle grades followed by formal methods in secondary school.

While designing separate courses may seem like a logical solution, many secondary mathematics preservice teachers pursue dual certification for middle and secondary mathematics. Informal methods must be taught in an authentic manner, particularly for those who will teach middle school students, without privileging formal approaches to inference. One possible approach to resolving this issue is to teach both formal and informal approaches in tandem. By working with raw data sets, preservice teachers could be asked to provide both a formal approach to generating a hypothesis and an informal approach through visual comparisons of data distributions. By contrasting informal and formal approaches, one a concrete representation and the other a symbolic representation, preservice teachers have the potential to develop a deeper and more connected understanding of inference as well as an understanding of the limitations and benefits associated with both approaches. Arguably, middle and secondary mathematics teachers should have an understanding of both approaches to inferential reasoning to be effective teachers.

Recommendations for Future Research

The findings of this study suggest the need for future research efforts. In this section, I advocate the value of additional research specific to: the development of

inferential reasoning, the characterization of formal inferential reasoning, the relationship between informal and formal inferential reasoning, and understanding how context and data display within tasks influence inferential reasoning.

The Development of Inferential Reasoning

A key finding of the study was a change from Unistructural to Multistructural inferential reasoning. Since the assessments were administered at the beginning and end of the course, this study is unable to conclusively report the degree of improvement associated with the first portion of the course focused on core and aggregate concepts versus the later portion specific to inferential content. The core concepts required to reason informally through the use of data displays were covered during the early portions of the statistics class with an emphasis on concepts and procedures. From the results of the midcourse interviews, a portion of the cohort demonstrated changes in their inferential reasoning by attending to additional aspects of the data in their responses. However, this study is unable to conclusively report precisely *when* the change in inferential reasoning occurred. It follows that additional research is required to carefully gauge the influence of instruction related to core and aggregate concepts on the development of inferential reasoning ability. The results of such research would provide guidance on the sequencing of statistical content and tasks to develop inferential reasoning and the role of informal and formal approaches to inference. In addition, studies that assess the influence of various course formats and curricular materials are needed to fully understand how inferential reasoning develops.

Formal Inference Characterization

This study addressed calls for additional investigations into the use of the SOLO taxonomy. The utility of the SOLO taxonomy in characterizing formal approaches to inference is promising given the results of this study, and warrants additional attention and refinement. Since justification was required after an inference was generated, adaptive reasoning was assessed primarily in relation to interpreting the result of the formal approach. One possible way to assess adaptive reasoning throughout formal inferential tasks is to require a prediction of the inference and a corresponding explanation of why the predication either aligned with the inference generated or did not. By requiring a prediction coupled with an explanation of results, a view is provided of how well the respondent understand the relationships between core and aggregate concepts in generating inferences. Developing robust descriptions of formal inferential reasoning at the four hierarchical reasoning levels based on student responses to formal inferential tasks is needed to provide a robust, cognitive framework for characterizing growth in this learning mode.

The Relationship between Informal and Formal Inferential Reasoning

In addition to refining characterizations of formal inferential reasoning, the relationship between informal and formal inference requires additional research. Statistics education leaders have placed their chips on a connection between the two, and this study supports that there is a relationship. However, informal inferential reasoning as a precursor to formal inferential reasoning has not yet been confirmed. In this study, a relationship between informal and formal inferential reasoning was identified through a correlational analysis. However, the number of formal responses in this analysis was

limited. Further research is needed to augment the nature of this relationship between informal and formal inferential reasoning in terms of the dependency of one reasoning approach on the other. From the results of this study, the SOLO taxonomy has shown to a viable approach to characterizing both types of approaches to inferential tasks and provides a framework for drawing comparisons.

Context and Data Representations of Tasks

One third of the cohort on each assessment seemed relatively unaffected by different contexts and data representations in their inferential reasoning responses. However, the majority of the cohort displayed variation in inferential reasoning across tasks on both assessments, and the context of tasks and choice of data displays may have played a role in the responses provided. Given the prominence of inferential reasoning in the middle and secondary school mathematics curriculum, an item bank of both informal and formal inferential reasoning tasks should be created for subsequent use with students of all ages as well as preservice mathematics teachers. Furthermore, responses to the items should be collected and analyzed so that the effects of context and data displays on inferential reasoning can be ascertained. The item test bank would also be a helpful tool for informing the instruction of middle and secondary students teachers as well as mathematics teacher educators.

Reflections

This study addressed an existing void of research focused on middle and secondary preservice teachers' knowledge of inferential reasoning after being afforded an opportunity to learn statistical content. With the increased role of statistics in the middle and secondary mathematics curriculum, an understanding of how well prepared teachers

are to teach statistics content in alignment with national standards is critical. Because inferential reasoning serves as a unifying topic in the statistics curriculum, a view into multiple aspects of the preservice teachers' knowledge was provided.

Upon reflection, certain aspects of the study were extremely helpful in both eliciting and characterizing the preservice teachers' inferential reasoning. The tasks selected as items on assessments were either derived from research studies or included in statistics education publications. Because justification and explanation was required to complete the tasks, robust responses were supplied and provided visibility into how the preservice teachers were thinking. Secondly, the SOLO taxonomy provided a means for comparing informal and formal inferential reasoning responses on an equivalent scale. Since this is the first research study to characterize both modes of inferential reasoning with the SOLO taxonomy, the results support Reading's (2007) recommendations to characterize informal and formal inferential reasoning in this manner.

Formal methods for inferential reasoning are often viewed as more cognitively advanced than informal approaches. From the findings of this research study, the possibility that students may advance their levels of reasoning in both informal and formal modes within similar timeframes offers an alternative perspective on how people learn to inferentially reason. This finding provides needed information about students' capacity to informally generate inferences and the associated capacity in formal approaches (Garfield & BenZvi, 2008). Integration of formal and informal approaches to inferential reasoning should set the stage for new curricular developments at the college level.

Lastly, national standards assert that reasoning and sense making should be the primary focus of secondary mathematics coursework. By classifying course content in terms of the emphasis placed adaptive reasoning, the degree that courses demand reasoning can be determined. The participants in this study are experiencing college courses that unfortunately do not reflect standards regarding reasoning. The need to alter teacher preparation programs to develop preservice teachers' ability to reason about and make-sense of statistical content is urgent given that their own experiences are at odds with how they will be asked to teach.

REFERENCES

American Statistical Association – National Council of Teachers of Mathematics. (1984). *The Statistics Teacher Network, 6,* 1-5.

Aquilonius, B. C. (2005). How do college students reason about hypothesis testing in introductory statistics courses? *Dissertation Abstracts International, 66*(02), 526. (UMI No. 3163105).

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

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

Biggs, J. B., & Collis, K. F. (1989). Toward's a model of school-based curriculum development and assessment using the SOLO Taxonomy. *Australian Journal of Education. 33*(2), 151-163.

Bright, G. W., Brewer, W., McClain, K., & Mooney, E. S. (2003). *Navigating through data analysis in grades 6-8.* Reston, VA: National Council of Teachers of Mathematics

Burrill, G., Franklin, C. A., Godbold, L, & Young, L. J. (2003). *Navigating through data analysis in grades 9-12.* Reston, VA: National Council of Teachers of Mathematics.

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

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

Cobb, P., McClain, K., & Gravemeier, K. P. E. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*(1), 1-78.

Conference Board of the Mathematical Sciences. (2001). *Issues in Mathematics Education Volume 11: The Mathematical Education of Teachers.* Providence: American Mathematical Society.

College Board. (2006). *College board standards for college success: Mathematics and statistics.* Mount Vernon, IL: Author.

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

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28(1)*, 96-105.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253-292.

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

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

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

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

Garfield, J. & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In J. Garfield & Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 397-409). Dordredcht, The Netherlands: Kluwer.

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M. Lovett and P. Shah (Ed.), *Thinking with Data (Proceedings of the 33rd Carnegie Symposium on Cognition)* (pp. 117-147). New York: Erlbaum.

Goel, V. (1992). A Comparison of Well-structured and Ill-structured Task Environments and Problem Spaces. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38*(5), 427-437.

Groth, R. E., & Bergner, J. A. (2005). Preservice elementary school teachers metaphors for the concept of statistical sample. *Statistics Education Research Journal, 4*(2), 27-42.

Heaton, R., & Mickelson, W. (2002). The learning and teaching of statistical investigation in teaching and teacher education. *Journal of Mathematics Teachers Education, 5,* 35-59.

Heid, M. K., Perkinson, D., Peters, S. A., & Fratto, C.L. (2005). Making and managing distinctions – the case of sampling distributions. In Lloyd, G. M., Wilson, M., Wilkins, J. L. M., & Behm, S. L. (Eds.). *Proceedings of the 27th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*.

Jones, D. L., & Tarr, J E. (2010). Recommendations for Statistics and Probability in School Mathematics Over the Past Century. In B. J. Reys, R. E. Reys, & R. Rubenstein (Eds.), *Mathematics Curriculum: Issues, Trends, and Future Direction: Seventy-second Yearbook* (pp. 65-76). Reston, VA: National Council of Teachers of Mathematics.

Jones, G. A., Langrall, C. W., Mooney, E. S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Dordredcht, The Netherlands: Kluwer.

Kilpatrick, K. J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children Learn mathematics*. New York: National Academy Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as a search for signals in noisy processes. *Journal for Research in Mathematics Education, 24*, 392-414.

Leathman, K. R., Lawrence, K., & Mewborn, D. S. (2005). Getting started with open-ended assessment. *Teaching Children Mathematics, 11*(8), 413-419.

Leavy, A. (2010). The challenge of preparing preservice teachers to teach informal inferential reasoning. *Statistics Education Research Journal, 9*(1), 46-67.

Liu, Y. & Thompson, P. (2005). Understandings of margin of error. In Lloyd, G. M., Wilson, M., Wilkins, J. L. M., & Behm, S. L. (Eds.). *Proceedings of the 27th annual meeting of the North American Chapter of the International Groups for the Psychology of Mathematics Education.*

Makar, K., & Confrey, J. (2004). Secondary teachers' reasoning about comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenges of developing statistical literacy, reasoning, and thinking* (pp. 353-374). Dordrecht, The Netherlands: Kluwer.

McClain, K. (2002). Supporting teachers' understanding of statistical data analysis: Learning trajectories as tools for change. *Proceedings of the 6th International Conference on Teaching Statistics.* Retrieved on April 16, 2011 from www.stat.auckland.ac.nz.

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal

reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction, 14*(2), 139-178.

Mendenhall, W., Beaver, B. & Beaver, R. (2005). *Introduction to Probability and Statistics.* Thomson Learning.

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*(1), 23-63.

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

Moore, D. (2004). *The basic practice of statistics* (3rd ed.). New York: W. H. Freeman.

Mrudulla, G., Scheaffer, R. L., Landwehr, J. M., Watkins, A. E., Barbella, P., Kepner, J., Newman, C. M., Obremski, T. E., & Swift, J. (1995). *Quantitative literacy series.* New York: Pearson Learning (Dale Seymour Publications).

National Council of Teachers of Mathematics. (2009). *Focus in High School Mathematics: Reasoning and Sense Making.* Reston, VA: author.

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

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: author.

National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010). *Common core state standards: Mathematics standards.* Retrieved July 17, 2010 from http://www.corestandards.org.

Nicholson, J. R., & Darnton, C. (2003). Mathematics teachers teaching statistics: What are the challenges for the classroom teachers? In *Proceedings of the 54th Session of the International Statistical Institute.* Voorburg, The Netherlands: International Statistical Institute.

Pratt, D., Johnston-Wilder, P. J., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistical Education Research Journal, 7*(2), 106-129.

Reading, C. (2007, August). *Cognitive development of reasoning about inference.* Discussant reaction presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution:

From a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68.

Rubin, A., & Rosebery, A. S. (1988, August). *Teachers' misunderstandings in statistical reasoning: Evidence from a field test of innovative materials.* Paper presented at the international Statistics Institute Round Table Conference, Budapest, Hungary, Voorburg, The Netherlands: International Statistics Institute.

Saldanha, L.A., & Thompson, P.W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257-270.

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

Stohl, H., & Tarr, J.E. (2002). Developing notions of inference with probability simulation tools. *Journal of Mathematical Behavior*, *21*(3), 319-337.

Tempelaar, D. T., Van der Loeff, S. S., & Gijselaers, W. H. (2007). A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal, 6(2),* 78-102.

Tversky, A. & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.) *Judgment under uncertainty: Heuristic and biases* (pp. 84-100). Cambridge, UK: Cambridge University Press.

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., & Verschaffel, L. (2005). Remedying secondary school students' illusion of linearity: Developing and evaluating a powerful learning environment. *Powerful environments for promoting deep conceptual and strategic learning, 2(7),* 115-132.

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

Watson, J. M. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal, 7*(2), 59-82.

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

Watson, J. M. (2001). Profiling teachers' competence and confidence to teach particular mathematics topics: The case of data and chance. *Journal of Mathematics Teacher Education, 4,* 305-337.

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

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145-168.

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

Williams, A.M. (1999). Novice students' conceptual knowledge of statistical hypothesis testing. In J.M. Truran, & K.M. Truran (Eds.), *Making the difference: Proceedings of the Twenty-second Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 554-560). Adelaide, South Australia: MERGA.

Zieffler, A., delMas, R., Garfield, J., & Gould, R. (2007, August). *Studying the development of college students' reasoning about statistical inference.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

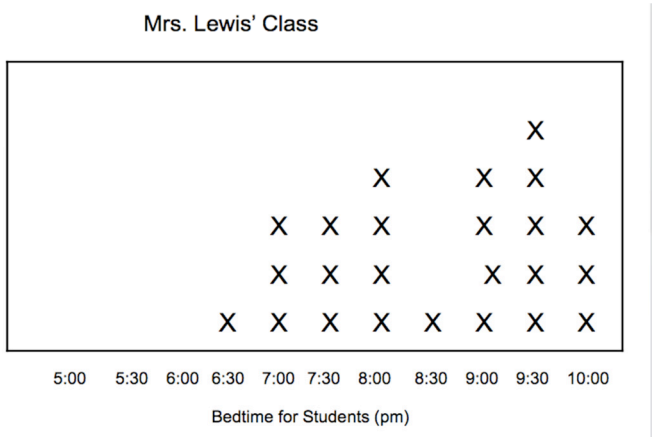
APPENDIX A: PRE-ASSESSMENT

Name:_____ Date:_____

Please attempt all tasks to the best of your ability and provide complete explanations where requested.

TASK 1: Middle School Student Bedtimes

Part A: The bedtimes of middle school students in Columbia were collected at the beginning of the school year. In the chart below are the data provided by one class. Please answer the following questions specific to this class.



A. What is the median bedtime value? _____

B. What does the median value represent or tell us about the data?_____

_____

C. What is the mode bedtime value? _____

D. What does the mode value represent or tell us about the data?_____

_____

E. What is the average bedtime value? _____

F. What does the average value represent or tell us about the data?_____

_____

G. What is the range for this set of data? _____

H. Are there any outliers (please list the values, if any)?_____

I. How would you describe the shape of the distribution of data?

_____

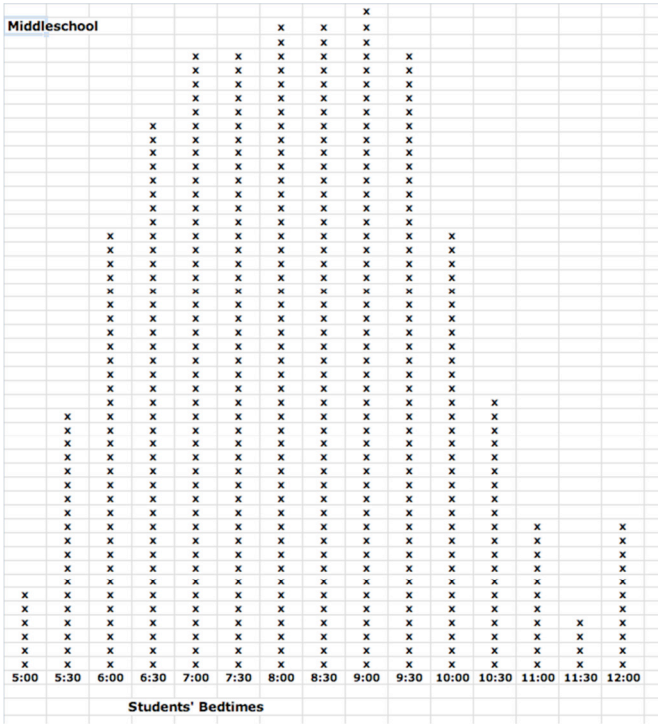
_____

J. How would you describe the bedtimes of this class to someone not looking at the data?

_____

_____

Part B: Now, the bedtimes of all the middle school students are shown in the chart below. Please answer the following questions specific to this set of data.



A. Without computing an exact answer, what would you estimate is the average bedtime for this large data set?

_____

B. Is the mean, mode or median the best indicator of the general bedtime for middle school students? Why?

_____

_____

.

C. How would you describe the shape of the distribution of this data set?

_____

_____

D. Compare the chart of Mrs. Lewis' class (in ***Part A***) to the chart of the entire school. Please comment on any notable differences? _____

_____

_____

_____

_____

E. What do you think might explain the differences you noticed?_____

_____

_____

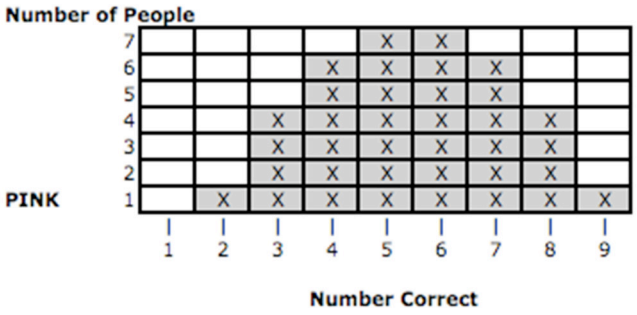
_____

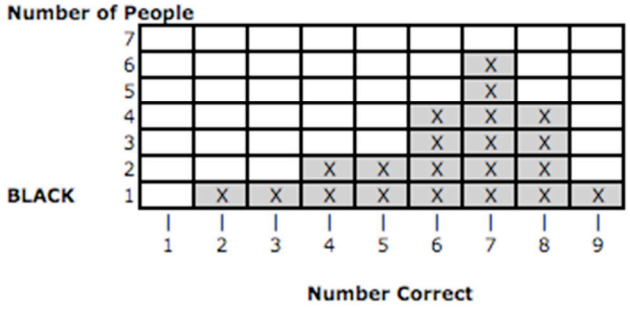
_____

_____

Task 2: Which Class Did Better?

Two classes are competing on quick recall math facts. One class is called the "Pink" class, and the other the "Black". The two classes both complete a quiz, and the results are shown below.





Which class did better? Please provide a complete explanation and any numerical information used in your rationale._____

_____

_____

_____

_____

_____

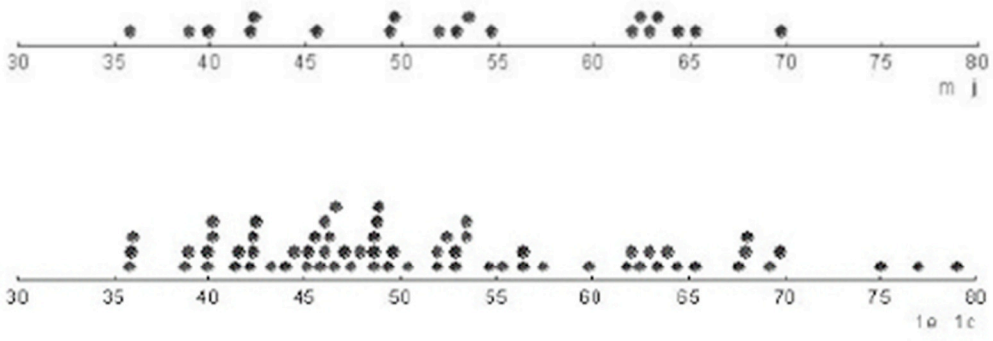
_____

_____

_____

Task 3: Weight of Grade 7 Students

Below are two sets of real data, the first one with 27 values and the second one with 67, showing the weights (in kilograms) of grade 7 students from Columbia, Missouri.



Question: Based on this actual data, what would you estimate the shape of the distribution of 1000 grade 7 weights might look like? Please sketch below and label your axes.

Provide an explanation and rationale for your sketch above. _____

_____

_____

_____

_____

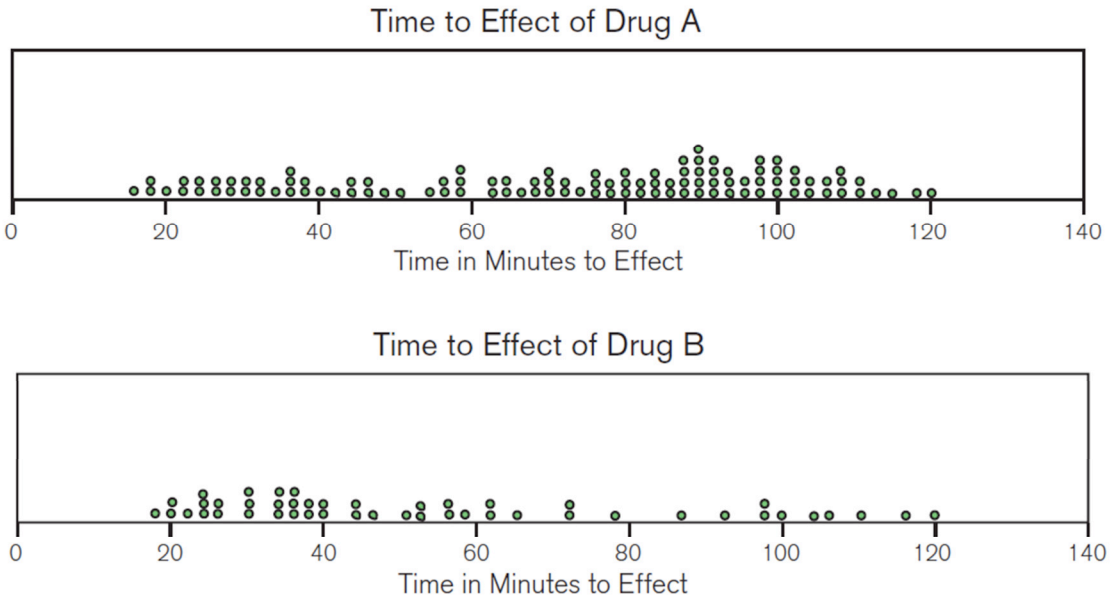
_____

_____

_____

Task 4: Which Treatment is More Effective?

Data is collected by two different pharmaceutical companies (Company A and Company B) on patients who suffer from migraine headaches. In both cases, the patients were told to take the medicine as soon as they experienced a headache and report how long it take to feel the relief effect of the medication. The results from each experiment are shown below.



Which medicine, Drug A or Drug B, would you recommend? Justify your choice below.

_____

_____

_____

_____

_____

_____

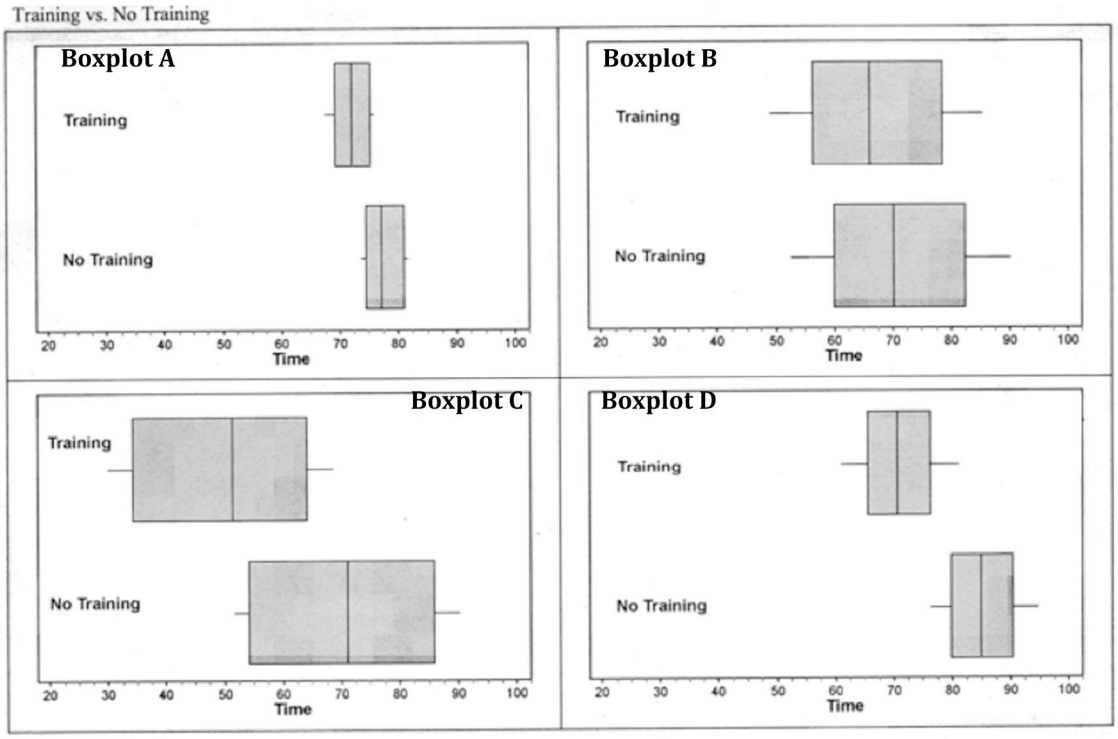
_____

_____

Task 5: Weight-Training Program Effectiveness for Track Athletes

Suppose there is a special summer camp for track athletes. There is one group of 100 athletes that run a particular race, and they are all pretty similar in their height, weight and strength. They are randomly assigned to one of two groups. One group has an additional weight-training program. The other group has the regular program without weight-training. All athletes from both groups run the race and their times are recorded, so that the data can be used to compare the effectiveness of the two training programs.

Below are four pairs of boxplots that compare the running times of athletes in the two different training programs (one with weight-training and one with no weight-training). Examine each pair of boxplots, and think about whether or not the sample data would lead you to believe that the difference in running times is caused by these two different programs. (Assume that everything else was the same for the students and this was a true, well-designed experiment).



A) Rank the four pairs of graphs on how convincing they are in terms of making an argument that the weight-training program was more effective in decreasing students' times from the least convincing to the most convincing evidence. Explain your reasoning.

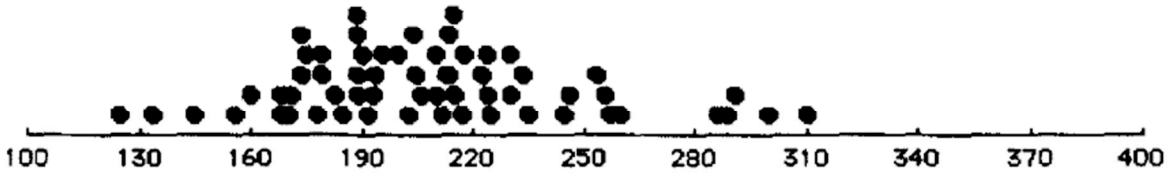
_____

_____

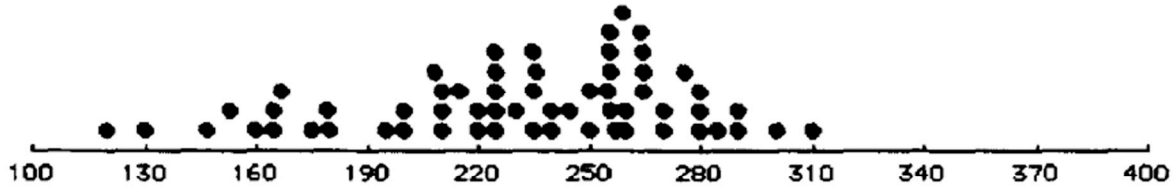
B) For the most convincing graph, would you be willing to generalize the effects of the training programs to all similar students on track teams based on these samples? Why or why not?

Task 6: **Diet and Cholesterol**

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



After Program



Before Program

Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

_____

_____

_____

_____

_____

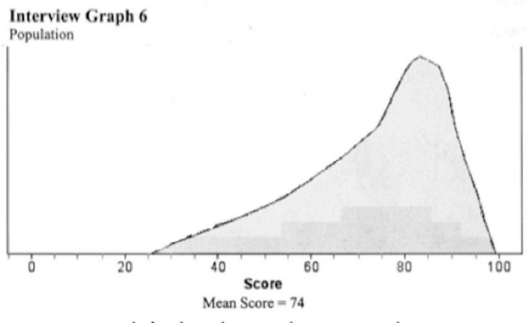
_____

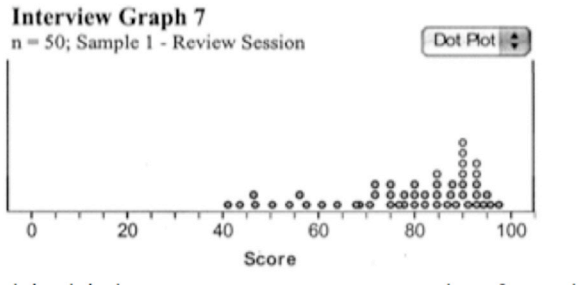
_____

_____

Task 7: Review Session Effectiveness

Below are two graphs of exam scores. The first one is a graph of exam scores representing many sections of students enrolled in an introduction to statistics course. For this population, the average score is 74. A random sample of 50 students in the class attended a review session with a teaching assistant prior to the exam. They were given the exact same exam as the other students in the population, but the mean exam score for these 50 students was 78, as shown in the second graph below.





A) Do you think that the teacher can attribute this higher average score to the fact that these students attended a review session? Explain.

_____

_____

_____

_____

B) Is there anything else that you need to know to help you decide that the higher sample mean was or was not due to chance? Explain.

_____

_____

_____

_____

_____

_____

_____

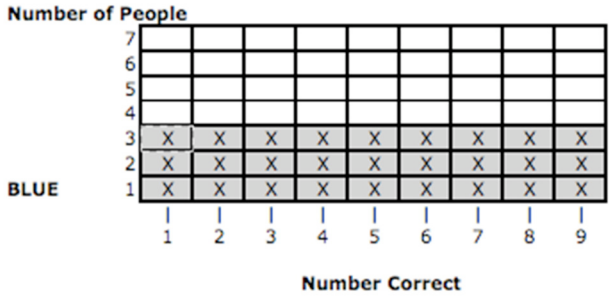
APPENDIX B: MIDCOURSE ASSESSMENT

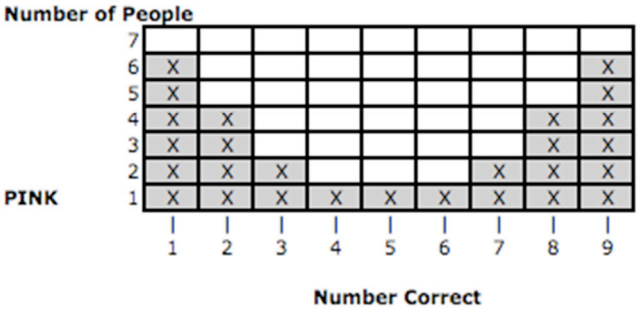
Name:_____ Date:_____

Please attempt all tasks to the best of your ability and provide complete explanations where requested.

Task 1: Which Class Did Better?

Two classes are competing on quick recall math facts. One class is called the "Blue" class with 27 students, and the other the "Pink" class also with 27 students. The two classes both complete a quiz, and the results are shown below.





Which class did better? Please provide a complete explanation and any numerical information used in your rationale.

_____

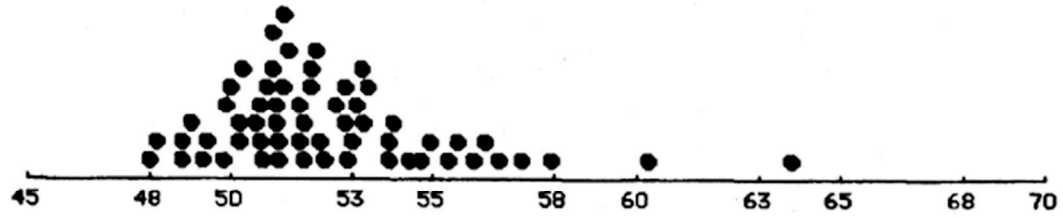
_____

_____

_____

Task 2: Speed Trap Effectiveness in Slowing Car Traffic

The city of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.



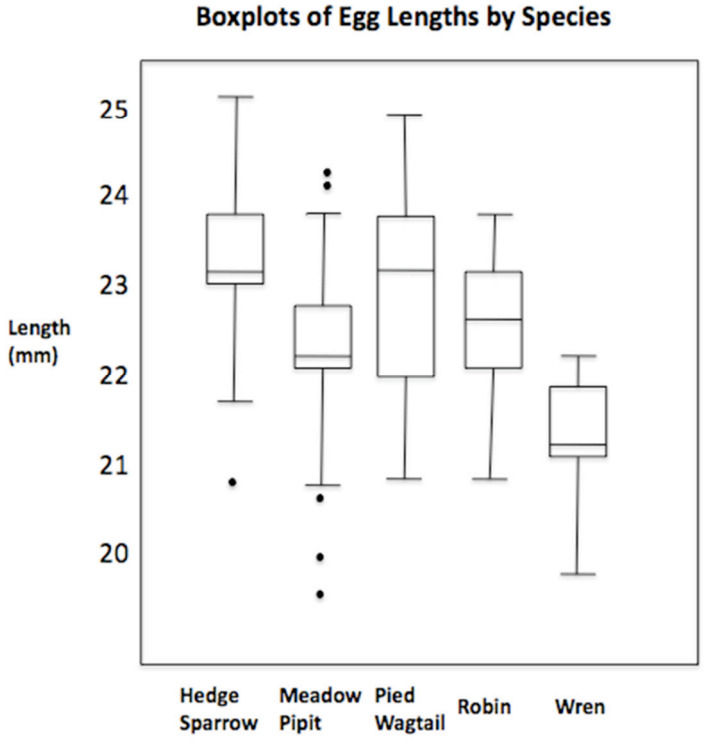
After Speed Trap

Before Speed Trap

Based on the data, was the speed trap effective in reducing the speed of traffic? Provide a detailed explanation for your position.

Task 3: Cuckoos Eggs

Cuckoos are known to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. These data give the lengths (mm) of cuckoo eggs found in the nests of other birds. A study investigates the difference in mean egg length (mm) of cuckoos' eggs according to the species of the foster parent.



With reference to the boxplot, do you think that there are any significant differences in mean egg lengths among the five species? Explain your reasoning.

_____

_____

_____

_____

_____

_____

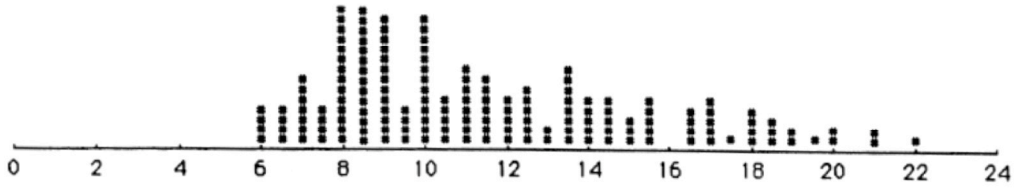
_____

_____

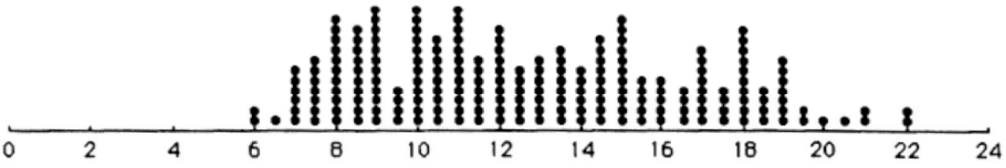
Task 4: Which Ambulance Service?

In St. Louis, the Clayton school district needs to select an ambulance service for emergencies that occur on school premises for the upcoming academic year. Two different ambulance companies provide service to the area: Acme and Lifetime.

Both companies provided the response times for emergency calls during the school year of 2009-2010 to other Clayton customers, as shown below. Acme provided data from 150 ambulance responses, and Lifetime provided data from 205 responses.



Acme



Lifetime

Based on the response times provided, which ambulance service would you recommend? Justify your choice below.

_____

_____

_____

_____

_____

_____

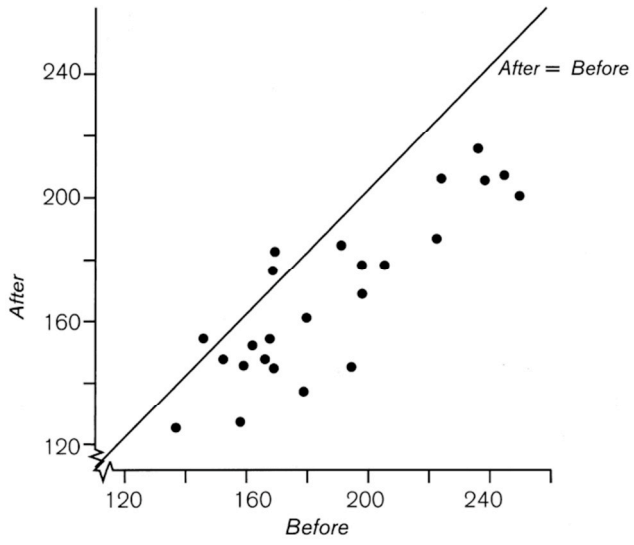
_____

Task 5: **Diet and Cholesterol**

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).

Cholesterol Levels

Before (mg/dL)	After (mg/dL)	Change (mg/dL)
195	146	49
145	155	-10
205	178	27
159	146	13
244	208	36
166	147	19
250	202	48
236	215	21
192	184	8
224	208	16
238	206	32
197	169	28
169	182	-13
158	127	31
151	149	2
197	178	19
180	161	19
222	187	35
168	176	-8
168	145	23
167	154	13
161	153	8
178	137	41
137	125	12



Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

Task 6: Pennies and Mints

Read the following scenario (Fong, Krantz & Nesbitt, 1986), and respond to the question posed.

> Joanna has a large collection of pennies with dates in the 1970's. Donny admires her collection and decides to start his own collection of pennies, but decides to collect only 1976 pennies because he wants to commemorate the Bicentennial. Looking through his pockets, he discovers he has only a dime. Examining it carefully, he finds that it is a 1971 dime, with a "D" (Denver) mint mark. Donny thinks it would be fun to collect 1976 pennies with the same initial as his name and asks Joanna what proportion of the 1976 pennies in her collection have a "D" mint mark on them.
>
> She doesn't know, but they decide to find out. They take the huge jar of her pennies out. Since the jar has thousands of pennies in it, Donny shakes the jar and then reaches into it and picks out a handful from the middle of the jar. Donny finds all the 1976 pennies that he scooped out (four of them) and finds that two of them have "D" mint marks. Because of this, he estimates that around 50% of all Joanna's 1976 pennies have the "D" mint mark.
>
> But Joanna looks through the other 36 pennies they have scooped out (dated 1970-1975 and 1977-1979) and discovers that only 2 of them have the "D" mint mark. She argues that only 4 of the 40 pennies altogether have the "D" mark, and estimates that around 10% of the 1976 pennies in her collection are "D" pennies.

Comment on the validity of Joanna's and Donny's reasoning. Whose conclusion about the 1976 pennies in Joanna's collection is more likely to be correct? Explain.

Task 7: Hiring of Managers and Discrimination

In 1972, 48 male bank supervisors were each randomly assigned a personnel file and asked to judge whether the person represented in the file should be recommended for promotion to a branch-manager job described as "routine" or whether the person's file should be held and other applicants interviewed.

The files were all identical except that half of the supervisors had files labeled "male" while the other half had files labeled "female". Of the 48 files reviewed, 35 were recommended for promotion. Twenty-one (21) of the 35 recommended files were labeled "male", and 14 were labeled "female."

Given the results of the simulations of randomly selecting files for promotion and the resulting number of "males" recommended for promotion, do you conclude that discrimination took place?

Explain why the simulated outcomes either support discrimination or not.

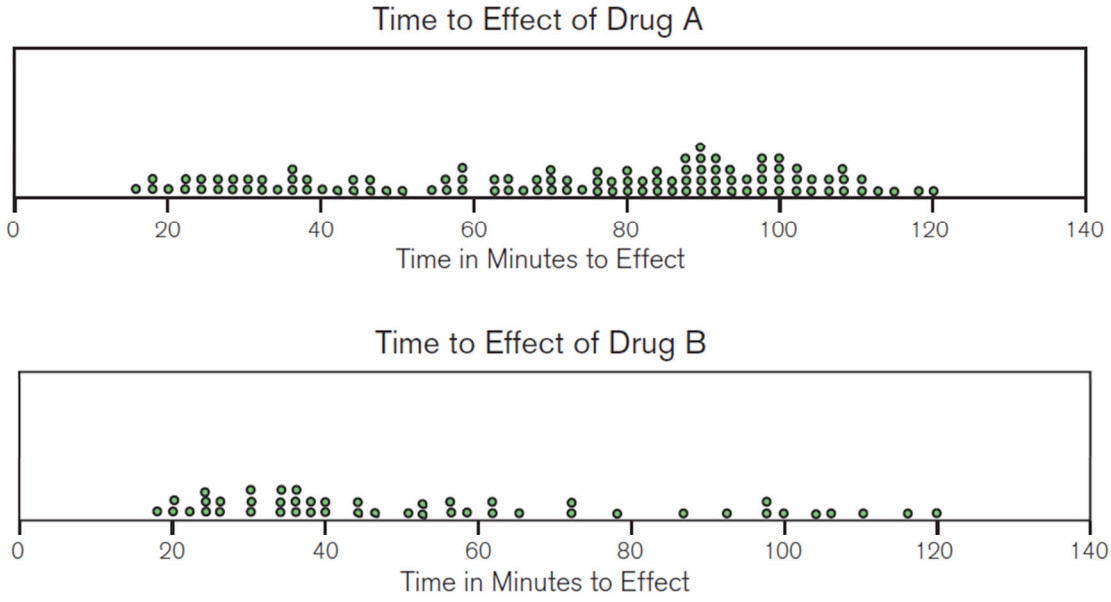
APPENDIX C: POST-ASSESSMENT

Name:_____ Date:_____

Please attempt all tasks to the best of your ability and provide complete explanations where requested.

Task 1: Which Treatment is More Effective?

Data is collected by two different pharmaceutical companies (Company A and Company B) on patients who suffer from migraine headaches. In both cases, the patients were told to take the medicine as soon as they experienced a headache and report how long it took until they felt the relief effect of the medication. The results from each experiment are shown below.



Which medicine, Drug A or Drug B, would you recommend? Justify your choice below.

_____

_____

_____

_____

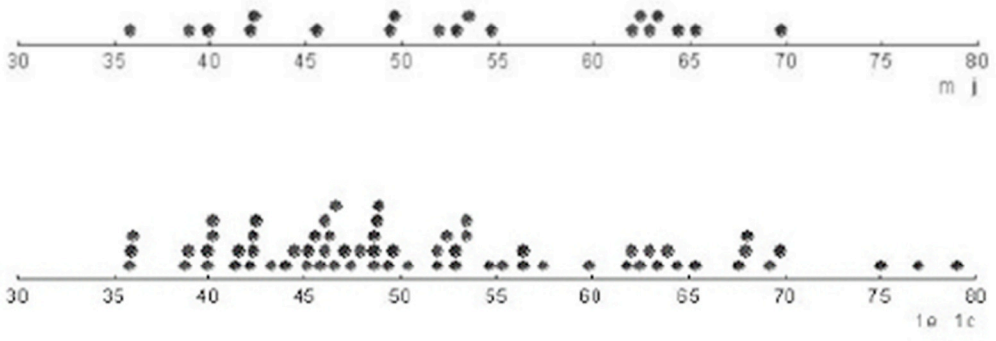
_____

_____

_____

Task 2: Weight of Grade 7 Students

Below are two sets of real data, the first one with 27 values and the second one with 67, showing the weights (in kilograms) of grade 7 students from Columbia, Missouri.



Question: Based on this actual data, what would you estimate the shape of the distribution of 1000 grade 7 weights might look like? Please sketch below and label your axes.

Provide an explanation and rationale for your sketch above.

_____

_____

_____

_____

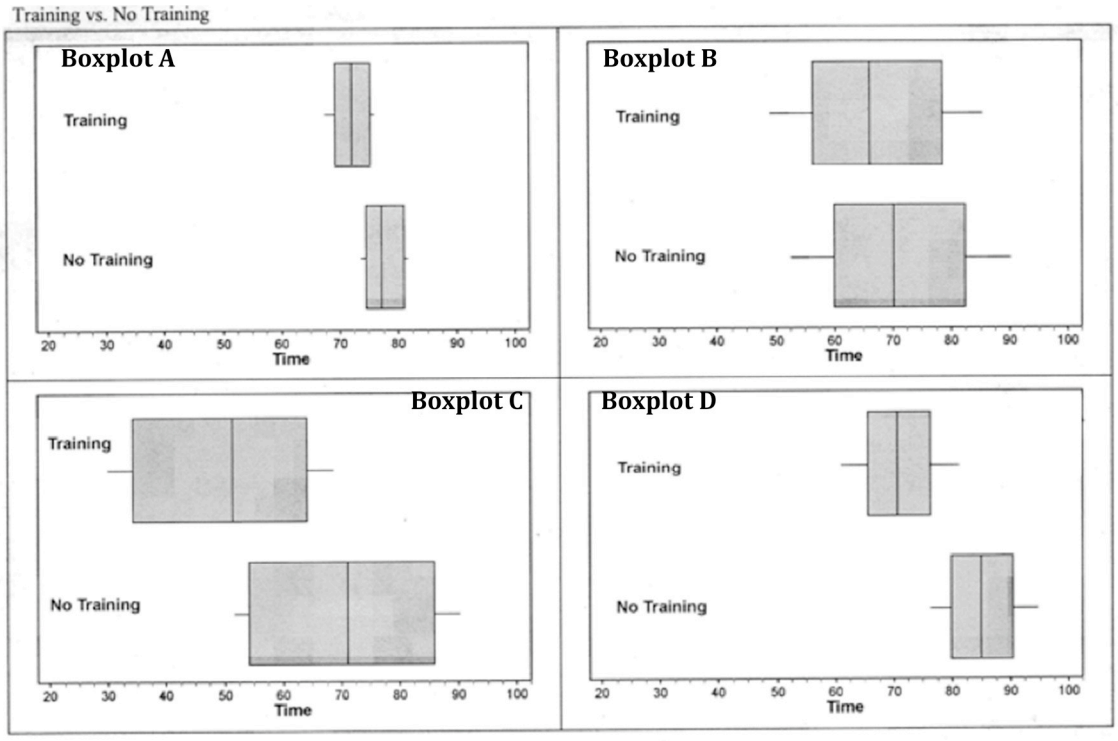
_____

_____

Task 3: Weight-Training Program Effectiveness for Track Athletes

Suppose there is a special summer camp for track athletes. There is one group of 100 athletes that run a particular race, and they are all pretty similar in their height, weight and strength. They are randomly assigned to one of two groups. One group has an additional weight-training program. The other group has the regular program without weight training. All athletes from both groups run the race and their times are recorded, so that the data can be used to compare the effectiveness of the two training programs.

Below are four pairs of box plots that compare the running times of athletes in the two different training programs (one with weight-training and one with no weight-training). Examine each pair of box plots, and think about whether or not the sample data would lead you to believe that the difference in running times is caused by these two different programs. (Assume that everything else was the same for the students and this was a true, well-designed experiment).



C) Rank the four pairs of graphs on how convincing they are in terms of making an argument that the weight-training program was more effective in decreasing students' times from the least convincing to the most convincing evidence. Explain your reasoning.

_____

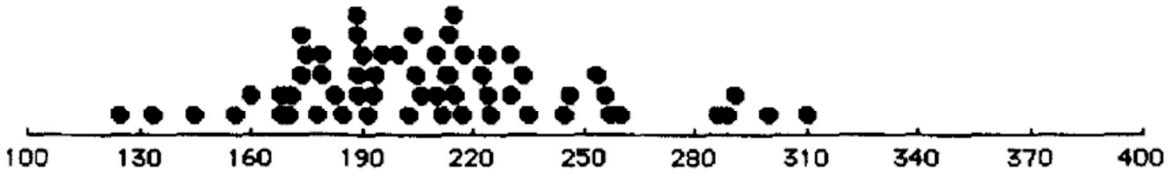
_____

_____

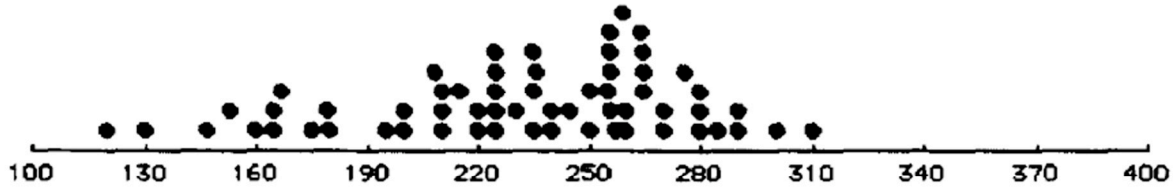
D) For the most convincing graph, would you be willing to generalize the effects of the training program to all similar students on track teams based on these samples? Why or why not?

Task 4: **Diet and Cholesterol**

High cholesterol is a contributor to heart disease. A study was conducted to investigate the effect of dietary change on cholesterol levels. Participants in the study voluntarily switched from a "standard American diet" to a vegetarian diet for one month. The data shown below are the participants' cholesterol levels before and after the dietary change, in milligrams of cholesterol per deciliter of blood (mg/dL).



After Program



Before Program

Assuming that lower levels of cholesterol are the goal, would you say that the change in diet is effective for lowering cholesterol or could similar results have been achieved by chance? Provide a detailed explanation below.

_____

_____

_____

_____

_____

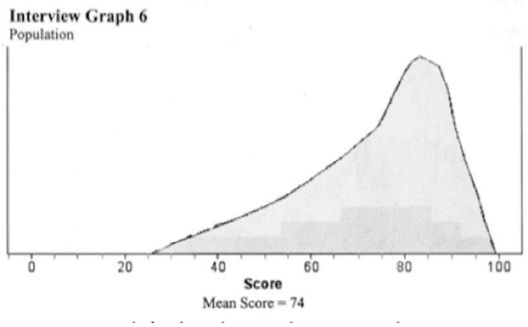
_____

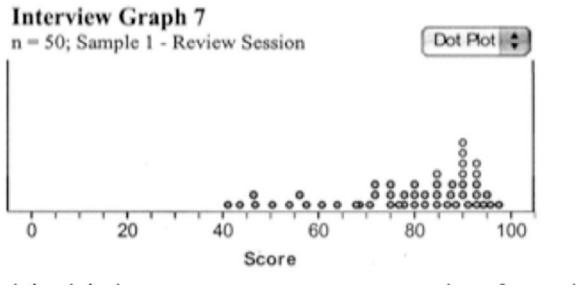
_____

_____

Task 5: Review Session Effectiveness

Below are two graphs of exam scores. The first one is a graph of exam scores representing many sections of students enrolled in an introduction to statistics course. For this population, the average score is 74. A random sample of 50 students in the class attended a review session with a teaching assistant prior to the exam. They were given the exact same exam as the other students in the population, but the mean exam score for these 50 students was 78, as shown in the second graph below.





B) Do you think that the teacher can attribute this higher average score to the fact that these students attended a review session? Explain.

_____

_____

_____

_____

C) Is there anything else that you need to know to help you decide that the higher sample mean was or was not due to chance? Explain.

_____

_____

_____

_____

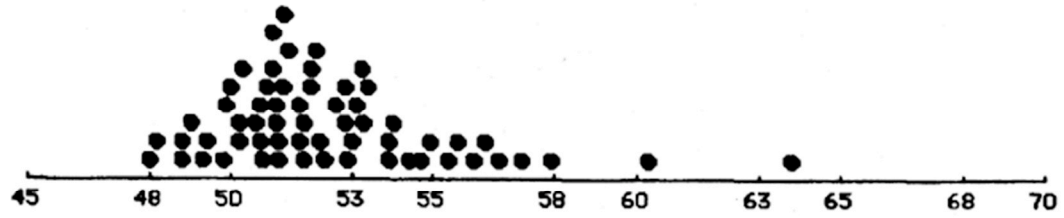
_____

_____

_____

Task 6: Speed Trap Effectiveness in Slowing Car Traffic

The city of Columbia introduced a police speed trap in a zone with a 50 mile per hour speed limit. The speeds of 60 cars are shown after the speed trap had been in place for some time and before.



After Speed Trap

Before Speed Trap

A) Before the speed trap was introduced, the average speed was 54.9 miles per hour for the 60 cars shown above, and 53.1 miles per hour afterwards. Similarly, the variance for these two samples was 12.5 squared mph before the speed trap and 7.34 squared mph afterwards.

Are the speeds of the cars significantly different after the speed trap was in place?

C) As a police officer, would you consider the speed trap to be successful or unsuccessful (practical significance)? Why?

Task 7: Hiring of Managers and Discrimination

In 1972, 48 bank supervisors were each randomly assigned a personnel file and asked to judge whether the person represented in the file should be recommended for promotion to a branch-manager job described as "routine" or whether the person's file should be held and other applicants interviewed.

The files were all identical except that half of the supervisors had files labeled "male" while the other half had files labeled "female". Of the 48 files reviewed, 35 were recommended for promotion. Twenty-one (21) of the 35 recommended files were labeled "male", and 14 were labeled "female."

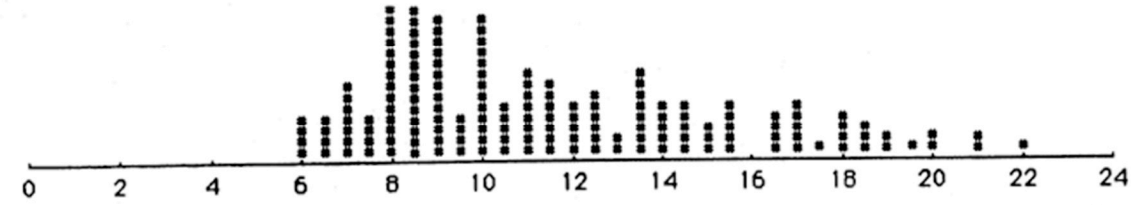
If the selection of the 35 candidates were purely fair in terms of gender given equal qualifications for promotion, we would expect that half the candidates would be male (17.5) with a standard deviation of 1.65 males.

Question: As a member of a jury, would you confidently support a verdict that the bank supervisors discriminated against female candidates? Support your response.

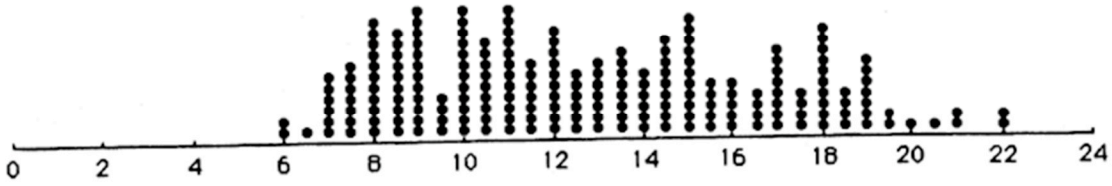
Task 8: Which Ambulance Service?

In St. Louis, the Clayton school district needs to select an ambulance service for emergencies that occur on school premises for the upcoming academic year. Two different ambulance companies provide service to the area: Acme and Lifetime.

Both companies provided the response times for emergency calls during the school year of 2009-2010 to other Clayton customers, as shown below. Acme provided data from 162 ambulance responses, and Lifetime provided data from 206 responses.



Acme



Lifetime

A) Based on the response times provided, $\bar{x}_{acme} = 11.6$ minutes and $\bar{x}_{lifetime} = 12.7$ minutes with samples variances, $\bar{s}^2_{acme} = 14.3$ and $\bar{s}^2_{lifetime} = 15.0$. Based on this information, is there a significant difference between the two ambulance services?

€ € € €

B) As a member of the school district, which ambulance service would you recommend? Why?

_____

_____

_____

_____

_____

_____

_____

_____

VITA

Maryann E. Huey was born on July 18, 1971 in Boulder, Colorado, where she lived until age 7. Her family moved to Houghton, Michigan, where Maryann completed high school in 1989 with a one-year move to Dunwoody, Georgia in grade 10. Maryann then attended the University of Michigan at Ann Arbor and earned a Bachelor's of Science in Electrical Engineering in 1993. While working for Sprint Corporation, Maryann completed a Masters in Business Administration from the Ohio State University at Columbus in 1996. After working for 10 years with Sprint Corporation, Maryann attended the University of Kansas at Lawrence and earned a Masters degree in Mathematics in 2006. Maryann taught mathematics at Rockhurst University for one semester and then moved with her family to Columbia, Missouri where she completed her Ph.D. in Mathematics Education from the University of Missouri at Columbia. Maryann has had to opportunity to teach college mathematics, mathematics methods and content courses for preservice teachers, and grade 5 mathematics content with the support of a master teacher.

Maryann is married to Brian Huey and has two children, Elise and Bryce. Maryann has accepted an Assistant Professor position at Drake University in Des Moines, Iowa in the Mathematics Department effective August 2011.