KNOWLEDGE AND UNDERSTANDING OF PROBABILITY AND STATISTICS TOPICS BY PRESERVICE PK-8 TEACHERS

A Dissertation

by

TAMARA ANTHONY CARTER

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Curriculum and Instruction

KNOWLEDGE AND UNDERSTANDING OF PROBABILITY AND STATISTICS TOPICS BY PRESERVICE PK-8 TEACHERS

A Dissertation

by

TAMARA ANTHONY CARTER

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Gerald O. Kulm
Committee Members,	Robert M. Capraro
	Victor L. Willson
	G. Donald Allen
Head of Department,	Dennie L. Smith

August 2005

Major Subject: Curriculum and Instruction

ABSTRACT

Knowledge and Understanding of Probability and Statistics Topics by Preservice PK-8 Teachers. (August 2005)

Tamara Anthony Carter, B.A., Rice University;

M.A., Rice University;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Gerald O. Kulm

Given the importance placed on probability and statistics in the PK-8 curriculum by the National Council of Teachers of Mathematics (2000) and on teachers by the Interstate New Teacher Assessment and Support Consortium (1995) and the Conference Board of the Mathematical Sciences (2001), it is important to know how well preservice teachers understand topics that are vital to a thorough understanding of the probability and statistics topics emphasized by national standards. It is necessary for a teacher to thoroughly understand the subject matter in order to teach effectively, but that is not sufficient. A teacher must also be able to successfully communicate with the students about that material. Therefore, this study utilized a standards- and literature-based assessment to study 210 preservice teachers with the goal of taking the first step in determining whether current PK-8 preservice teachers are prepared to teach select probability and statistics topics specified in standards documents. The assessment contains 11 probability and statistics items with a total of 23 parts in a variety of short-answer, multiple-choice, and extended-response formats. It is described in detail in

Chapter III and reproduced in Appendix A.

A confirmatory factor analysis indicated that for this sample of PK-8 preservice teachers, the assessment measured the underlying constructs on which it was based. Preservice teachers' ability to answer these items varied greatly. For short-answer and multiple-choice items, the percentage of preservice teachers incorrectly answering an item was as high as 87% and as low as 18%. For extended-response items, incorrect answers were provided by as few as 12% of the participants on one item and by as many as 83% on another. Individual responses were analyzed to illustrate correct conceptions and misconceptions of these preservice teachers. There was not a statistically significant difference between responses based on the grade band the participants were preparing to teach, but students specializing in mathematics and science did perform better than other participants. Although effect sizes were small, the amount of time elapsed since an elementary statistics class was taken and the number of methods courses taken were positively associated with performance on this assessment.

DEDICATION

To my wonderful husband, my terrific parents, and all the family, friends, teachers, and mentors who have supported me over the years – THANK YOU!

ACKNOWLEDGEMENTS

I would like to thank my biological and academic families and my friends for all their help over the years. Dr. Kulm, thank you for your constant guidance throughout this degree and for allowing me to work so closely with the Middle School Mathematics Project. Dr. Robert Capraro, thank you for consistently helping me to learn and grow as a mathematics educator and for those perfectly placed words of encouragement. Dr. Willson, thank you for guiding me all the way from *t*-tests to Structural Equation Modeling as you helped me develop as an educational researcher. Dr. Allen, thank you for keeping me grounded in the *mathematics* of mathematics education. I would also like to thank Dr. Mary Margaret Capraro, Dr. Bob Hall, and (soon to be Dr.) Susan Cromwell Duncan for the important role that you played throughout this dissertation. Your feedback and cooperation from the very first stage of development of the pilot study through to the completion of this dissertation have been very helpful. I would also like to thank those who provided feedback through the many iterations of the assessment. To all the instructors who helped me solicit participants for this study, thank you. My heartfelt gratitude goes to the participants for taking the time out of their busy schedules to help me with this study. Linda, Robert, Adam, Mary Margaret, Judy, and Chris, I learned so much from our collaboration on manuscripts and presentations and enjoyed the time we spent working together - thank you. I would also like to thank my fellow graduate students. I enjoyed exchanging ideas and learning about education and life from one another.

Of course, my education did not start with this degree. I would like to thank all my teachers from Ms. Tetrick in Kindergarten, to Ms. Kubala and Ms. Atkins in elementary, to Mr. Cole, Ms. Sturdivant, and Mr. Percifield in junior high, to Mr. Waters and Mr. K. in high school, to Dr. Tapia and Dr. Papakonstantinou at Rice. Every one of you taught me something special, but I owe a special debt of gratitude to Terry Don Waters, my high-school mathematics teacher and mathematics team coach. Thank you for the many hours you spent with us preparing for competitions, for all those Saturdays you dedicated to us, and for solidifying my love of mathematics and teaching.

Although education is important, some of the most important teachers are not confined to school buildings. I would also like to thank all my family and friends who loved me enough to pretend to that it was acceptable when I felt I needed to work on a paper rather than visit. I appreciate your support and understanding. To my parents, Mark and Gail Anthony, thank you for your constant love and support. You have always been there helping me to reach my potential. Thanks to Cecil and Nelda Carter. When I married your son, I also gained a new set of parents who have loved and encouraged me. Thank you to all my “extra parents” back home and to the Hosman family who have always been there to make my small family feel large. To my grandfather, Mark Anthony, thanks for living life to the fullest and reminding me to do the same. To my puppies, thank you for all the hours that you spent at my feet keeping me company while I pursued this degree. Most of all, I would like to thank my husband, Tracy Carter. Thank you for your patience and day-to-day support and love. You believed in me, encouraged me to pursue my dreams, and helped me every step of the way. I love you!

TABLE OF CONTENTS

Page

CHAPTER Page

Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

An emphasis has been placed on data analysis and probability for grades pre-kindergarten (PK) through 12 in the last 15 years by organizations such as the National Council of Teachers of Mathematics (NCTM; 1989, 2000), the Interstate New Teacher Assessment and Support Consortium (INTASC; 1995), and the Conference Board of the Mathematical Sciences (CBMS; 2001). Therefore, teacher preparation programs need to ensure that their graduates are prepared to meet the challenges of teaching this curriculum. Around the time that these standards were enacted, Shaughnessy (1992) reported that teachers were not adequately prepared to teach this strand of the curriculum. However, the emphasis on statistical concepts and skills in the K-12 curriculum began to appear around the time that the majority of current preservice teachers started Kindergarten, so they should be more prepared on these topics than previous generations of teachers.

Other than the work of Jane Watson and her colleagues (c.f., Torok & Watson, 2000; Watson, 2001; Watson & Mortiz, 2000), very little recent research has focused on the probabilistic and statistical conceptions of preservice teachers. Tversky and Kahneman (1974, 1983) laid the groundwork for the study of misconceptions of chance. Many studies have followed their lead, but few have focused on preservice teachers. Similarly, research on the use of measures of center and spread is abundant, but a focus on preservice teachers' knowledge of these topics is lacking. Shaughnessy (1992)

This dissertation follows the style of *Journal for Research in Mathematics Education.*

pointed out that “we have to first deal with teachers’ misconceptions before we can expect them to be competent at helping their students to overcome misconceptions" (p. 484). In order to deal with teachers’ misconceptions, we must first understand the nature of those misconceptions. In addition to the problem of misconceptions, O'Connell (1999) reported “many college students seem to grasp only a partial understanding of fundamental concepts and procedures in probability” (p. 3). Partial understandings are not acceptable for preservice teachers. They must know the correct answers and also *why* those answers are correct (Schulman, 1986). Additionally, Schulman distinguished between subject-matter knowledge (similar to a content expert) and pedagogical content knowledge (the understanding of how the topics are understood – or misunderstood – by people) stressing that both are of vital importance (1986). In order to improve probability and statistics education for student in grades K-12, we must ensure that teachers are prepared with the background probability and statistics knowledge and the understanding of how to teach it.

Statement of the Problem

Given the importance placed on probability and statistics in the PK-8 curriculum (NCTM, 2000), it is important to know how well preservice teachers understand topics that are vital to a thorough understanding of the probability and statistics topics emphasized by national standards for the grade bands they expect to teach. It is necessary for a teacher to thoroughly understand the subject matter in order to teach effectively, but that is not sufficient. A teacher must also be able to successfully communicate with the students about that material. Therefore, this study utilized a

standards- and literature-based assessment containing 11 probability and statistics items with a total of 23 parts in a variety of short-answer, multiple-choice, and extended-response formats to study 210 preservice teachers with the goal of taking the first step in determining whether current PK-8 preservice teachers are prepared to teach select probability and statistics topics specified in standards documents. Specifically, this assessment examined preservice teachers' knowledge and understanding of measures of center and measures of spread. Utilizing multiple-choice and extended-response items, the study also examined the possible presence of some probabilistic misconceptions that have been well documented among children and the general population but have been relatively unexplored among preservice teachers.

Research Questions

This study took the first step in determining how well PK-8 preservice teachers are prepared to teach certain probability and statistics skills and concepts specified in national standards. The goal of the study was to examine the nature of the knowledge and understanding of probability and statistics topics by preservice PK-8 teachers who had already taken or were currently enrolled in elementary statistics. Specifically, the following questions were explored.

1. To what extent are PK-8 preservice teachers successful at answering multiple-choice and short-answer questions concerning measures of center, measures of spread, and common misconceptions of chance?

2. What is the quality of the written explanations of PK-8 preservice teachers to questions concerning measures of center, measures of spread, and common misconceptions of chance?
3. What is the nature of the difference between PK-8 preservice teachers' ability to determine a correct answer and their ability to explain that answer for a sample of probability and statistics questions concerning measures of center, measures of spread, and common misconceptions of chance?
4. What is the nature of the difference between preservice EC-4, 4-8 Language Arts and Social Studies, and 4-8 Mathematics and Science teachers on correctness of responses and quality of explanations for a sample of probability and statistics questions concerning measures of center, measures of spread, and common misconceptions of chance?
5. What is the effect of cumulative course exposure to probability and statistics topics on PK-8 preservice teachers' ability to choose a correct answer and their ability to explain that answer for a sample of probability and statistics questions concerning measures of center, measures of spread, and common misconceptions of chance?

CHAPTER II

BACKGROUND LITERATURE

Probability and statistics have a daily impact on people's lives. People make decisions based on risk assessments (either a formal or intuitive form of probability), and people analyze statistical information in order to digest information from media sources. Although probability and statistics have been a part of the university curriculum for over a century, their presence in K-12 curricula has been minimal (Truran, 2001). Probability and statistics entered the mainstream of western world curricula in the 1960s with the "new mathematics" curricula (Truran, 2001) and received renewed emphasis beginning in the 1980's by organizations such as the National Council of Teachers of Mathematics (NCTM), the Interstate New Teacher Assessment and Support Consortium (INTASC), the American Statistical Association (ASA), and other mathematically oriented organizations that comprise the Conference Board of the Mathematical Sciences (CBMS).

Teaching and Learning Standards

Studies investigating teacher's preparedness to teach probability and statistics (Shaughnessy, 1992) revealed that teachers were not prepared to teach these topics. In 2001, over a decade after NCTM's *Curriculum and Evaluation Standards* (1989) emphasized probability and statistics for PK-12 students, the CBMS (2001) study revealed that teachers were least prepared to teach the probability and statistics strand of the curriculum.

According to the NCTM standards, students in grades PK-2 are expected to organize and describe data and discuss events in terms of likely and unlikely; students in grades 3-5 are expected to represent data using bar graphs, use appropriate measures of center, describe the distribution of data, and classify events as likely or unlikely; additionally, students in grades 6-8 are expected to create and use histograms and boxplots, interpret measures of center and spread, and compute probabilities for simple compound events (NCTM, 2000). Of course, teachers need to know what their students are expected to learn, but they must know much more.

The INTASC standards were designed to assist individual states in the creation of licensure examinations by identifying specific concepts that all new teachers should know. According to the INTASC standards, all teachers should understand (1) the concept of "data representation to describe data distributions, central tendency, and variance through appropriate use of graphs, tables, and summary statistics"; (2) "analysis and interpretation of data, including summarizing data, and making or evaluating arguments, predictions, recommendations, or decisions based on an analysis of the data"; and (3) "probability as a way to describe chance or risk in simple and compound events" (INTASC, 1995, p. 20-21). Specific examples identified within these standards include bar graphs, mean, median, mode, range, standard deviation, and "knowing that when a fair coin is tossed the fraction of tosses that are heads approaches ½ as the number of flips increases" (INTASC, 1995, p. 21).

The Conference Board of the Mathematical Sciences (CBMS; 2001) also published recommendations for teachers. Among other things, CBMS (2001) proposed

that elementary teachers should have experience with describing data including shape, center, and spread and understanding probability including randomness, judgments under uncertainty and likelihood measurements. Their recommendations for middle school teachers expanded these notions to emphasize spread and variability and further use of probability.

Analysis of the NCTM, INTASC, and CBMS standards reveals many commonalities. The statistics topics are primarily focused on the study of summary data, specifically central tendency and variability. The probability topics center on the likelihood of events primarily using the concepts of randomness, conjunction, and sample size (e.g., implications of the law of large numbers).

Statistics

"For those who have traditionally been left out of the political process, probably no skill is more important to acquire in the battle for equity than statistical literacy" (Konold & Higgins, 2003, p. 193), yet statistics causes confusion for many. Moore (1990) stresses that the goal of including statistics in the curriculum is to promote thinking, interpreting, and use of judgment rather than to learn a specific set of computational skills. In order for this goal to be achieved, the teachers must have a deep understanding of the content (CBMS, 2001; Heaton & Mickelson, 2002) that is free of misconceptions so they can support activities that are meaningful and significant in the classroom.

Measures of Center

To many people, the word “average” is synonymous with the statistical term “arithmetic mean”. However, median and mode are also appropriate measures for finding the center of a set of data. According to Watson & Moritz (2000), all three measures of central tendency (i.e., measures of center) have been covered on the collegiate level for over a century, and mean has been used at the pre-colligate level for most of that time, but median and mode were not a standard part of the PK-12 curriculum prior to the 1989 NCTM standards. Measures of center are a fundamental building block necessary for the understanding of statistical inference (Konold & Higgins, 2003) which is often the focus of elementary statistics classes. Yet according to the research of Pollatsek, Lima, and Well (1981), students, even at the collegiate level, view at least one measure of center, the mean, computationally rather than conceptually. Similarly, Rubin and Rosebery (1988) found that both students and their teacher were confused that the median could remain unaltered when additional data points were added to a set. A thorough understanding of all three measures of center is vital before meaningful comparisons can be made among them.

Variability of Data

Understanding of variation is essential for placing measures of center in context. In fact, Moore lists “the omnipresence of variation in processes” as one of the “core element of statistical thinking” (1990, p. 135). Many people think of standard deviation when variability is mentioned because this is the measure stressed in college statistics classes. However, measures of variability such as the range and interquartile range are

also useful and more appropriate for younger students. Unfortunately, very little research has been conducted on students' understanding of variability (Watson, Kelly, Callingham, & Shaughnessy, 2003) even though national standards formally include variability beginning in middle school and utilize variability concepts in lower grades (NCTM, 2000).

Probabilistic Misconceptions

According to Konold (1995), students enter statistics courses with incorrect intuitions that are extremely difficult to alter. Probability is one of the mathematical topics for which misconceptions are highly likely (Shaughnessy, 1981). In order to equip teachers to identify and remediate probabilistic misconceptions in their students, the teachers' probabilistic misconceptions must first be confronted (Shaughnessy, 1992).

What is a misconception? According to *The American Heritage Dictionary* (2000), a misconception is "a mistaken thought, idea, or notion". However, the following research on probabilistic misconceptions utilizes the psychological notion in which a misconception is not merely a careless error, but is systematic in nature. Misconceptions are rarely eliminated by simply teaching the topics of class; misconceptions must be directly confronted and remediated (Byrnes, 2001; Carpenter & Hiebert, 1992; Mevarech, 1983). Unless misconceptions are directly confronted, both "the misconceptions and the scientific principles may coexist as separate islands of knowledge" (Carpenter & Hiebert, 1992, p. 89). Amos Tversky and Daniel Kahneman worked in the fields of psychology and cognitive science. Their work on risk assessment showed that people hold consistent beliefs about probability that directly contradict

established probability theory. The work of Tversky and Kahneman laid the foundation for future studies of probabilistic misconceptions.

Representativeness Heuristic

Tversky and Kahneman's (1974) work on judgments under uncertainty led them to discover that the same heuristics that help people make daily decisions also lead to "severe and systematic errors" (p. 1124). One of these heuristics is the representativeness heuristic in which people use the degree to which event A *resembles* class B to estimate the *probability* that event A belongs to class B (Tversky & Kahneman, 1974).

One of the problems with this heuristic is that it does not take base-rate frequencies into consideration. In one study, Tversky and Kahneman provided participants with a stereotypical description of an engineer and the information that this person was randomly drawn from a group containing 30 lawyers and 70 engineers. The participants were asked to provide the probability that the person described was an engineer. In general, participants based their answers on the description without attention to the base rate. When a similar question was provided that used a description that could fit a lawyer equally as well as an engineer, participants still ignored the base rate and provided a 0.5 probability that the person was an engineer. However, if no description was provided, participants accurately gave 0.7 as the probability that the person described was an engineer (Tversky & Kahneman, 1974). This pattern of responses indicates that descriptions, even non-informative ones, trigger the use of the representativeness heuristic rather than formal probabilistic knowledge. Davidson (1995)

presented second, fourth, and sixth graders with a similar problem using descriptions pertaining to elderly citizens and found that less than a quarter of these students used the base-rate information to answer the question.

Another misconception that Tversky and Kahneman placed under the heading of representativeness is one pertaining to the randomness of chance. “People expect that a sequence of events generated by a random process will represent the essential characteristics of that process even when the sequence is short” (Tversky & Kahneman, 1974, p. 1125). For example, the sequence of coin tosses THTHHT is perceived to be more random (and thus more likely) than HHHTTT. Similarly, the sequence HHHTTT is perceived to be more likely than HHHHTH because HHHTTT appears to represent the fairness of the coin better (Tversky & Kahneman, 1974). However, although more of the 64 outcomes have three heads than five heads, each of the 64 outcomes for the tossing of six coins is equally likely. Fast (1997) presented a similar question to student teachers preparing to teach secondary mathematics and found that a third of them failed to give the correct answer.

Recognizing the pervasive belief that all samples are representative of the population regardless of sample size, Tversky and Kahneman framed a study with a problem asking whether a large hospital or a small hospital would be more likely to record more days in which at least 60% of the babies born were male. Results show that the participants did not take sample size into consideration and erroneously answered that both hospitals were equally likely to have such days (Tversky & Kahneman, 1974). However, use of the law of large numbers specifies that the large hospital should be

much more representative of the population (even split of boys and girls) than the small hospital, so the small hospital is much more likely to deviate from the even split and have days in which at least 60% of the births are boys. Fischbein and Schnarch (1997) presented the same problem to 500 students from fifth grade through college. Over 55% of their participants indicated that the probabilities were equivalent in both hospitals, and only 1% correctly indicated that the small hospital would be more likely to have a large deviation from half boys and half girls (Fischbein & Schnarch, 1997). The results of their study also suggest that the attainment of additional mathematical concepts (such as ratios and proportions) can confound students' use of the law of large numbers (Fischbein & Schnarch, 1997).

Conjunction Fallacy

The probability of the conjunction of two events *A* and *B* (i.e., $P(A \cap B)$, sometimes written as $P(A \& B)$, which is the probability that both *A* and *B* will happen) is less likely than $P(A)$ and than $P(B)$. This is because an item has to fit both qualifications (*A* and *B*) in order to qualify for $P(A \& B)$. However, Tversky and Kahneman (1983) investigated people's conception of the relative rankings of $P(A)$, $P(B)$, $P(A \& B)$ and found that they often deviated from the probabilistic law of $P(A \& B) \leq \min(P(A), P(B))$ which means that they fell prey to the conjunction fallacy. When conjunction problems are placed in a social context, it is reasonable to believe that people are judging the representativeness of the situation rather than the actual probabilities (Tversky & Kahneman, 1983). However, Gavanski and Roskos-Ewoldsen (1991) utilized prompts with a less social context and found that participants still used

the conjunction fallacy, but did so at a lesser rate. Use of the conjunction fallacy has been reported for students ranging from second grade to college (Davidson, 1995; Fischbein & Schnarch, 1997). Conjunction problems differ from most mathematical concepts in that people answer incorrectly (invoke the conjunction fallacy) more often when problems are presented in a concrete form then in an abstract form (Tversky & Kahneman, 1983).

CHAPTER III

METHODOLOGY

The intent of this study was to assess the nature of the knowledge and understanding of probability and statistics topics by preservice PK-8 teachers. This study utilized a within-stage mixed-model design (cf. Johnson & Onwuegbuzie, 2004) where initially the data were analyzed quantitatively to investigate the overall implications followed by a qualitative analysis. In order to understand the nature and structure of the participants' responses, the quantitative results were used to inform the selection of a purposeful sample for qualitative analysis. Through constant comparison, similarities in responses were identified and unifying commonalities were grouped into meta-categories (Denzin & Lincoln, 2000).

Instrumentation

The primary data collection tool was the participant assessment. The instructor survey was used to gather information about possible differences between class sections on the administration of the participant assessment or material presented in class.

Participant Assessment

The assessment instrument used for this study was a compilation of items modified from a review of the literature. This assessment was subjected to a validation study and a four-stage pilot test (Mertens, 2005). The validation study relied on the informed opinions of eight experts in the design or revision of assessment items or the content under investigation. The four-stage pilot study was an iterative process of data gathering and item revision. In the first stage, the assessment was administered to three

informed practitioners and content experts and revised. In the second stage, the assessment was administered to nine informed practitioners or content experts and revised again. In the third stage, the assessment was administered to a sample of 35, primarily undergraduate, students. The data were analyzed and the items were revised in consultation with three methodological and content experts. In the fourth stage, the assessment was administered to a sample of 88 students enrolled in an undergraduate elementary statistics class. The data were analyzed and the items were revised in consultation with two informed practitioners, four content experts, and a methodological expert into the form used for this study. The assessment contains eleven probability and statistics items, some of which contain multiple parts, and six demographic items. Four versions of this assessment were created (see Appendix A for version A of the assessment). Two versions have the probability items first. The other two have the statistics items first. Within each of these sets, one assessment has items in the original order and one has the items in reverse order.

Theoretical model

The participant assessment was framed by the theoretical model in Figure 1. This model was based on the standards set forth by the National Council of Teachers of Mathematics (NCTM), the Interstate New Teacher Assessment and Support Consortium (INTASC), and other mathematically oriented organizations that comprise the Conference Board of the Mathematical Sciences (CBMS).

The items are divided into two parts, Likelihood and Summary Data, which are the latent variables in the theoretical model. The compound measured variables

Randomness, Law of Large Numbers, and Conjunction are representations of the latent construct Likelihood. Central Tendency and Variability are the two compound measured variables representing the latent construct Summary Data.

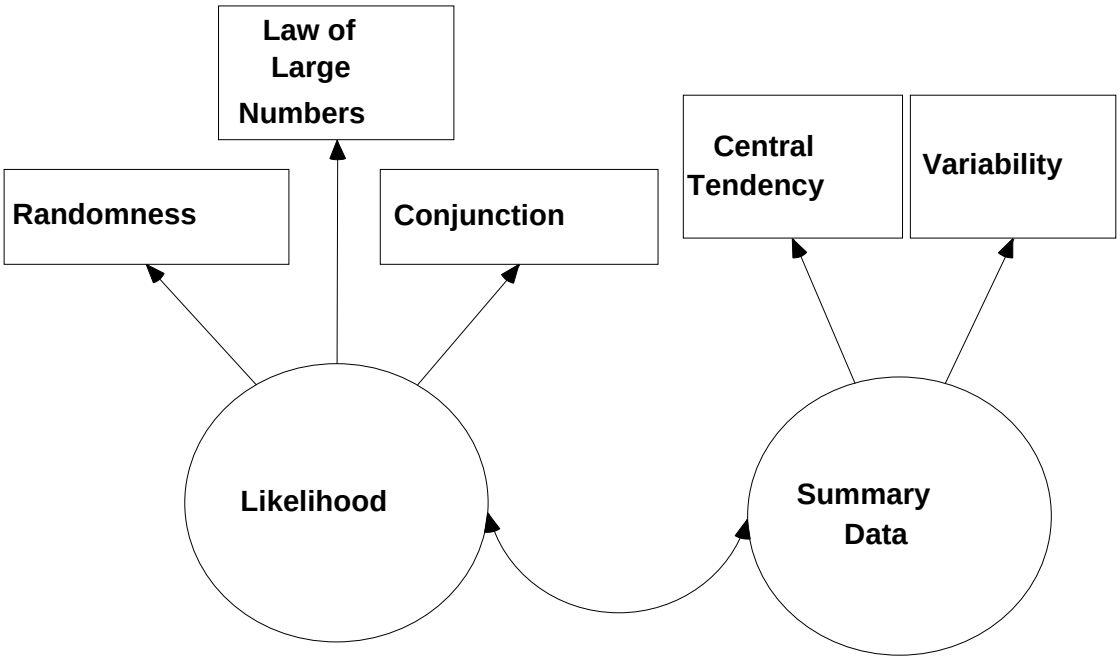


Figure 1. Theoretical Model for the Study.

Item and assessment etymology

The items comprising the Randomness variable are 12, 13a, and 13b from version A (see Appendix A). The idea for item 12 originated in the literature on the gambler's fallacy (Tversky & Kahneman, 1974), an example used in the INTASC standards (1995) for "probability as a way to describe chance or risk in simple and compound events", and Fast's (1997) WDYTTCA (What Do You Think The Chances

Are?) instrument. Item 13a was inspired by the literature on the representativeness heuristic (Tversky & Kahneman, 1974). The style of the item aligned more closely with Fast (1997) but encompassed more of the known misconceptions as distracters. Item 13b was written to elicit reasoning about item 13a and the participants' pedagogical content knowledge per Schulman's definition as "subject matter knowledge *for teaching*" (1986, p. 9).

Items 11a, 11b, 14a, and 14b combined to form the Law of Large Numbers variable. Items 11a and 11b utilized the law of large numbers reasoning in reverse (as *n* decreases). Item 11a required participants to know the effect of sample size on the arithmetic mean. Item 11b measured participants' concept of the effect of sample size on standard deviation of the data. The basis for item 14a was Tversky and Kahneman's (1974) classic Hospital Problem assessing the belief that even a small sample is representative of the population, but the item was rephrased into a teaching situation. Item 14b was designed so that participants would display their reasoning about item 14a and their pedagogical content knowledge.

Compound Conjunction was measured through items 15, 16, 17a, and 17b. Items 15 and 17a focused on the conjunction fallacy (Tversky & Kahneman, 1974). Item 15 was intended to be devoid of context in contrast to item 17a that employed a social context. Pedagogical content knowledge and knowledge of probabilistic conjunctions was required for item 17b. Item 16 was a conjunction item of a computational nature similar to those commonly found in an introductory probability and statistics class.

Central Tendency was measured by item 7 (which was recoded into 7a, 7b, and 7c to reflect knowledge of arithmetic mean, median, and mode, respectively), 10a, 10b, 10c, and 10d. Item 7 addressed arithmetic mean, median, and mode from a more creative standpoint than simply computing those measures from given data. Item 10 was modified from Mokros and Russell's (1995) typicality problem into a teaching setting requiring knowledge of all three measures of center.

Items 8, 9b, 11c, and 11d were used to measure Variability. Item 8 related to the use of a standard deviation to describe percentages of data that fall within a certain range. Item 9b measured participants' concept of range as a measure of variability. Although item 9a is not part of the conceptual model, it was used as background information to help identify if incorrect answers to 9b could be attributed to difficulties reading graphs. Item 11c was intended to elicit a conceptual definition for standard deviation, and item 11d was intended to elicit information about the percentage of data encompassed within one, two, or three standard deviations of the mean under a normal curve. However, participant responses did not align with the expected outcomes for the items (i.e., some participants answered item 11c in 11d or in both 11c and 11d or item 11d in 11c). Therefore, for scoring purposes, items 11c and 11d were scored together and reported as 11e.

Instructor Survey

The instructor survey was used to gather information about possible differences between courses or sections on the administration of the participant assessment or class preparation (see Appendix B). Instructors were asked how many students were in the

class so that a response rate for participants could be determined. Instructors were asked about communications with the students concerning the assessment and incentives given for taking the assessment in order to account for differences in the analysis of the data and discussion of the results. The instructors were asked to evaluate each of the items on the participant assessment as a measure of content validity.

Data Collection

Both instruments, the instructor survey and the participant assessment, were created for distribution using Educational Survey Tool (Beser, 2004), which allowed for online data collection. From the time that the instruments were posted until they were removed, participants and instructors had 24-hour access. Therefore, instruments could be completed at the convenience of the participants and instructors. A link to the consent form was on the first page of each instrument (Appendixes C and D) and was available to be printed. No data was collected if the participant did not agree to the consent form.

Participant Recruitment and Assessment Administration

This study occurred at a large, southern, public university. The population for the study was preservice teachers pursing certification for grade bands PK-4 or 4-8 who were enrolled in or had previously taken an introductory statistics class. Course descriptions were examined, and seven courses were identified because their intended student body matched the population for this study. These seven courses were STAT 303 Statistical Methods, taught through the Department of Statistics, EPSY 435 Educational Statistics, taught through the Department of Educational Psychology, and five courses taught through the Department of Teaching, Learning, and Culture: ECFB 440 Math

Methods in Early Childhood Education, MEFB 450 Social Studies Methods in the Middle Grades, MEFB 460 Math Methods in the Middle Grades, MASC 450 Integrated Mathematics, and MASC351 Problem Solving (see Appendix E for course descriptions). All instructors of these courses were asked to participate (seven for STAT 303, three for EPSY 435, three for ECFB 440, two for MEFB 450, one for MEFB 460, two for MASC351, and all three instructors for the single section of MASC 450). One of the MEFB 450 instructors did not respond to inquiries. According to degree plan design and confirmed with the instructor, all of the students enrolled in MEFB 460 were also enrolled in MASC 450, so these students were accessed through MASC 450. All other instructors agreed to ask their students to access the assessment online between November 8th and November 24th, 2004. The Educational Survey Tool (EST) randomly assigned one of the four versions of the assessment to each student upon access to the primary website. Students were allowed to use a calculator but no other aids on the assessment. Approximately half of the instructors provided no incentives for the students to complete the assessment, and the other instructors offered minimal extra credit (see Appendix F). Most of the 17 participating instructors told the class about the assessment and either emailed reminders and/or posted the link on the class website. However, three instructors did not use class time to mention the assessment, and one did not use electronic communication. Using this procedure, approximately 900 students were asked to respond to the assessment. The computer logged 359 responses, and comparison of names provided by the participants indicated that these came from 346 different

participants. The response rate was approximately 38% (see Appendix F for more details).

According to the names provided by the participants on their assessments, seven of the participants stopped the assessment during the demographic items and took the assessment in full at a later time. Therefore, the eight surveys stopped during the demographics (one participant did this twice) were removed from the sample. Two participants who filled out the survey twice submitted answers that were almost identical to their previously submitted answers, so the two assessments were combined into a single set of data for each participant. Two participants who resubmitted responses had answers of vastly different quality in the submissions. Therefore, all assessments (two for one participant and three for the other) were removed from the sample. This left 344 assessments each from a different participant. One participant spent less than 10 minutes answering the assessment and provided only single-phrase answers to the extended-response items, so these data were considered not to be valid and were removed from the sample. The responses from another participant appeared to have been truncated by a computer malfunction. Therefore, these data were removed from the sample. The responses from 342 participants were analyzed. Of these participants, 223 (65.2%) were enrolled in a program leading to teachers' certification. Of these participants, 213 were either currently enrolled in or had already completed one of the two elementary statistics courses listed on the degree plans for these certification routes. Three of these participants were seeking certification in grades 8-12, 41 were seeking certification in grades 4-8 specializing in Mathematics and Science, 59 were seeking certification in

grades 4-8 specializing in Language Arts and Social Studies, and 110 were seeking certification in grades EC-4. The 210 participants seeking certification in grade EC-4 or 4-8 match the career goal for the population desired for this study and comprise the sample investigated for this study.

Of the 210 participants in the sample, 203 (96.7%) were female. None of these participants were freshmen, 9 (4.3%) were sophomores, 34 (16.2%) were juniors, 166 (79.0%) were seniors, none were graduate students, and 1 (0.5%) was not classified in any of these ways. Of the participants, 9 (4.3%) were Hispanic/Latino, 196 (93.3%) were White (non-Hispanic), 2 (1.0%) were African Americans, none were Asian, Pacific Islander, American Indians, or Alaskan Natives, and 3 (1.4%) chose not to identify themselves with any of these categories. Of the participants, 93 (44.3%) were enrolled in elementary statistics at the time of the assessment.

Instructor Survey

The instructor survey was posted on November 28th, 2004 and remained available until the last instructor completed the survey. Periodic reminders were sent to the instructors requesting the completion of this survey until the response rate was 100%.

Data Coding and Reliability

Extended responses were analyzed using the process of constant comparison (Lincoln & Guba, 1985). The categories that emerged from the comparisons were ranked according to the correctness of the answer on a 6-point rubric of (5) desired answer, (4) acceptable answer, (3) incomplete answer, (2) incorrect answer, (1) restatement of the

question, and (0) no answer (see Appendix G). Consistency of the coding is essential to the usefulness of the coding results, so intra-rater and inter-rater reliability studies were conducted (Huck, 2004; Mertens, 2005). To measure the intra-rater reliability of the rubric coding, a random sample of 100 item responses were selected and re-analyzed. Intra-rater reliability was 87%. To establish the inter-rater reliability of the rubrics, the 100 item random sample was re-analyzed by a second rater who is a mathematics education expert with experience teaching PK-8 students and PK-8 preservice teachers. Inter-rater reliability was 83%.

To facilitate data analysis, the answer choices for each multiple-choice item were ranked from most correct to least correct. This hierarchical ranking was based on the severity of the errors that would logically lead to each carefully constructed distracter. Similarly, possible answers for short-answer items were categorized and ranked from most correct to least correct. For consistency of scores throughout the assessment, the hierarchy for multiple-choice and short-answer items was translated into a 6-point scale with a five representing the best answer. Many of these items required fewer than six levels of scoring. To aid in the interpretation of the results and to allow the scale to more closely approximate interval scaling, answers choices were placed along the scale according to the severity of the error leading to that answer rather than blindly utilizing consecutive or evenly spaced rubric scores (see Appendix G). This system facilitated the use of statistical tests requiring ordinal data, allowed for the interpretation of tests requiring interval data, and maintained a consistent meaning of a desired answer throughout the assessment.

Data Analysis

Because four versions of the assessment were administered, an analysis of variance was conducted to determine if performances on the different versions were similar enough for the data to be aggregated. Additionally, the theoretical model utilized in the construction of the assessment was analyzed using structural equation models.

The question concerning the extent to which preservice PK-8 teachers were successful at answering multiple-choice and short-answer items concerning measures of center, measures of spread, and common misconceptions of chance was answered using descriptive statistics (including mean, median, mode, standard deviation, and frequencies of responses for each rubric level).

The question concerning the quality of written explanations of PK-8 preservice teachers to items concerning measures of center, measures of spread, and common misconceptions of chance, was answered using descriptive statistics and representative participant responses from the data analysis based on content and pedagogical content knowledge literature. The categories that emerged from the analysis of the written explanations were used to identify structuring schemata underlying intuitions.

The question concerning the nature of the difference between PK-8 preservice teachers' ability to choose a correct answer and their ability to explain that answer for a sample of probability and statistics items was explored using a correlation and a *t*-test. For items that utilized both a non-extended-response (multiple-choice or short-answer) portion and an extended-response portion, a composite score was attained for each portion by summing rubric scores. Therefore, both composite scores were scaled using

the same metric. A Pearson *r* correlation was used to estimate the relationship between the composite scores for non-extended-response and extended-response items. Additionally, a dependent samples *t*-test was used to compare the means of these two composite scores. Finally, corresponding non-parametric analyses were conducted.

The question concerning the nature of the difference between preservice EC-4, 4-8 Language Arts and Social Studies, and 4-8 Mathematics and Science teachers on correctness of responses and quality of explanations for a sample of probability and statistics items concerning measures of center, measures of spread, and common misconceptions of chance, was explored using ANOVAs followed by planned contrasts. Two comparisons were of interest: comparing performance on the two latent factors in the theoretical model, Likelihood and Summary Data, and comparing performance on the two types of items utilized in the assessment, extended-response and non-extended-response. Both of these comparisons were initially explored using one-way ANOVAs (4 total), then planned contrasts were used to explore where the differences occurred (2 contrasts for each ANOVA). Finally, corresponding non-parametric analyses were conducted.

The question concerning the effect of cumulative course exposure to probability and statistics topics on PK-8 preservice teachers' ability to determine a correct answer and their ability to explain that answer for items concerning likelihood and summary data, was explored using structural equation models. These models were run using the maximum likelihood method and evaluated using a variety of fit indices (Thompson, 2004). Although the χ^2 test of statistical significance is heavily dependent upon sample

size, it measures the difference between the covariance matrix produced by the sample data and one produced based on model constraints and is useful for comparing nested models (Thompson, 2004). Because lack of significance of the χ^2 test indicates a good fit of the data to the model, lower values for χ^2 within nested models indicates a better fit. Root-mean-square error of approximation (RMSEA) assesses the error between the model fit and the population covariances, so values less than .06 are indicative of a good fit (Thompson, 2004). The Comparative Fit Index (CFI) compares the model to a baseline model that assumes all measured variables are uncorrelated, so CFI values greater than or equal to .95 typically indicate a good fit of the data to the model (Thompson, 2004). Because these three indices measure different aspects of model fit, all three were reported. This question was further investigated with a multivariate regression using cumulative scores for each content type (likelihood and summary data) as dependent variables and time since statistics, a variable derived from the number of semesters since the participant took elementary statistics, and the number of pedagogically oriented courses that the participant has taken as independent variables. Finally, another multivariate regression was conducted using the same dependent variables with extended response and non-extended response as the independent variables.

CHAPTER IV

ANALYSIS

To investigate the performance of preservice PK-8 teachers, the data were restricted to the 210 participants who were enrolled in or had previously completed elementary statistics and who indicated they were pursuing certification in Early Childhood Education, Grades 4-8 Language Arts and Social Studies, or Grades 4-8 Mathematics and Science. All variables were coded on an ordinal scale with a maximum score of five. Skewness and kurtosis were computed for each measured variable and for the total assessment score, which was computed by summing the rubric scores on all items in the theoretical model (see Table 1). Differences in the total assessment score of participants across different versions of the assessment were investigated to determine if all the scores could be investigated as one group or if differences existed by versions. Due to the data's approximation to interval scaling, the lack of extreme skewness and kurtosis (Hopkins & Weeks, 1990) for the total assessment score (see Table 1), the fit of the total assessment scores to a normal distribution (a Kolmogorov-Smirnov Test yielded $z = 0.92$ with $N = 210$, $p = .364$), and homogeneity of variances for assessment versions (Levene's statistic = 1.94, $p = .125$), the assumptions for use of a one-way analysis of variance (ANOVA) were met. The ANOVA testing differences on total score between the four versions of the assessment indicates that there is not a statistically significant ($\alpha = .05$) difference among the test versions, $F(3, 206) = 1.74$, $p = .160$, $\eta_p^2 = .025$, so the data from all 210 participants

Table 1
Skewness and Kurtosis

	Skewness	Ratio of Skewness to Standard Error of Skewness	Kurtosis	Ratio of Kurtosis to Standard Error of Kurtosis
F7a Mean	-0.64	-3.84	-1.60	-4.79
F7b Median	-1.89	-11.28	1.75	5.24
F7c Mode	-2.00	-11.94	2.45	7.34
F8 Std MC	-1.90	-11.34	2.06	6.16
F9a Graph	-1.82	-10.86	1.45	4.35
F9b Range	-0.96	-5.71	-0.88	-2.62
F10a Mode	-1.84	-10.96	3.99	11.94
F10b Median	0.09	0.55	-0.27	-0.81
F10c Mean	-0.31	-1.84	-0.97	-2.92
F10d Typical	0.15	0.88	1.90	5.69
F10e Typical Type	-1.02	-6.06	-0.43	-1.29
F11a Std Mean	-1.74	-10.39	1.39	4.16
F11b Std Dev	1.11	6.63	-0.13	-0.39
F11c Std Exp	-0.65	-3.86	0.63	1.88
F11d Std Dist	-0.34	-2.02	-0.28	-0.84
F11e STD	-1.08	-6.41	0.98	2.93
F12 Coin HT	-2.55	-15.18	4.74	14.19
F13a Birth	-1.97	-11.74	4.16	12.46
F13b Birth	-0.15	-0.91	-0.63	-1.89
F14a Coin Party	0.03	0.17	0.26	0.77
F14b Coin Party	1.29	7.68	1.67	4.98
F15 Dice	0.70	4.19	-1.12	-3.36
F16 Coin HH	-0.40	-2.41	-1.48	-4.43
F17a Doc	0.37	2.23	-1.66	-4.96
F17b Doc	0.91	5.41	-0.12	-0.35
Total Score	-0.26	-1.54	0.27	0.80

were aggregated for analysis. Because the data approximate interval scaling but are ordinal in nature by strict definition, non-parametric statistics will be provided throughout this study to assure the reader that the results reported are reflective of the data rather than the assumption of approximate interval scaling. In keeping with that

goal, a Kruskal-Wallis one-way analysis of variance by ranks of total assessment score by assessment version also led to the conclusion that the scores did not differ by assessment version, $\chi^2(3, N = 210) = 5.53$, $p = .137$.

Analysis of the Theoretical Model

The theoretical model derived from the literature (Figure 1 on page 16) was the basis for the assessment used in this study. Before research questions were analyzed based on this model, the fit of the data to the model had to be explored. The measured variables (Randomness, Law of Large Numbers, Conjunction, Central Tendency, and Variability) in the model represent composite variables formed by summing the rubric scores for the associated items. Randomness was composed of items 12, 13a, and 13b. Law of Large Numbers was composed of 11a, 11b, 14a, and 14b. Items 15, 16, 17a, and 17b comprised Conjunction. Central Tendency was formed from items 7a, 7b, 7c, 10a, 10b, 10c, and 10d. Variability was formed by items 8, 9b, and 11e (item 11e was the rubric score used to combine the responses from 11c and 11d into one item).

The first step in exploring this model was to analyze the measurement model in which the assessment items were the measured variables, and the measured variables from the theoretical model (Randomness, Law of Large Numbers, Conjunction, Central Tendency, and Variability) were the latent variables. All structural equation models in this study utilized the maximum likelihood method of estimation. The resulting model is displayed in Figure 2, the path coefficients for the variables are listed in Table 2, and the correlations between latent variables are listed in Table 3.

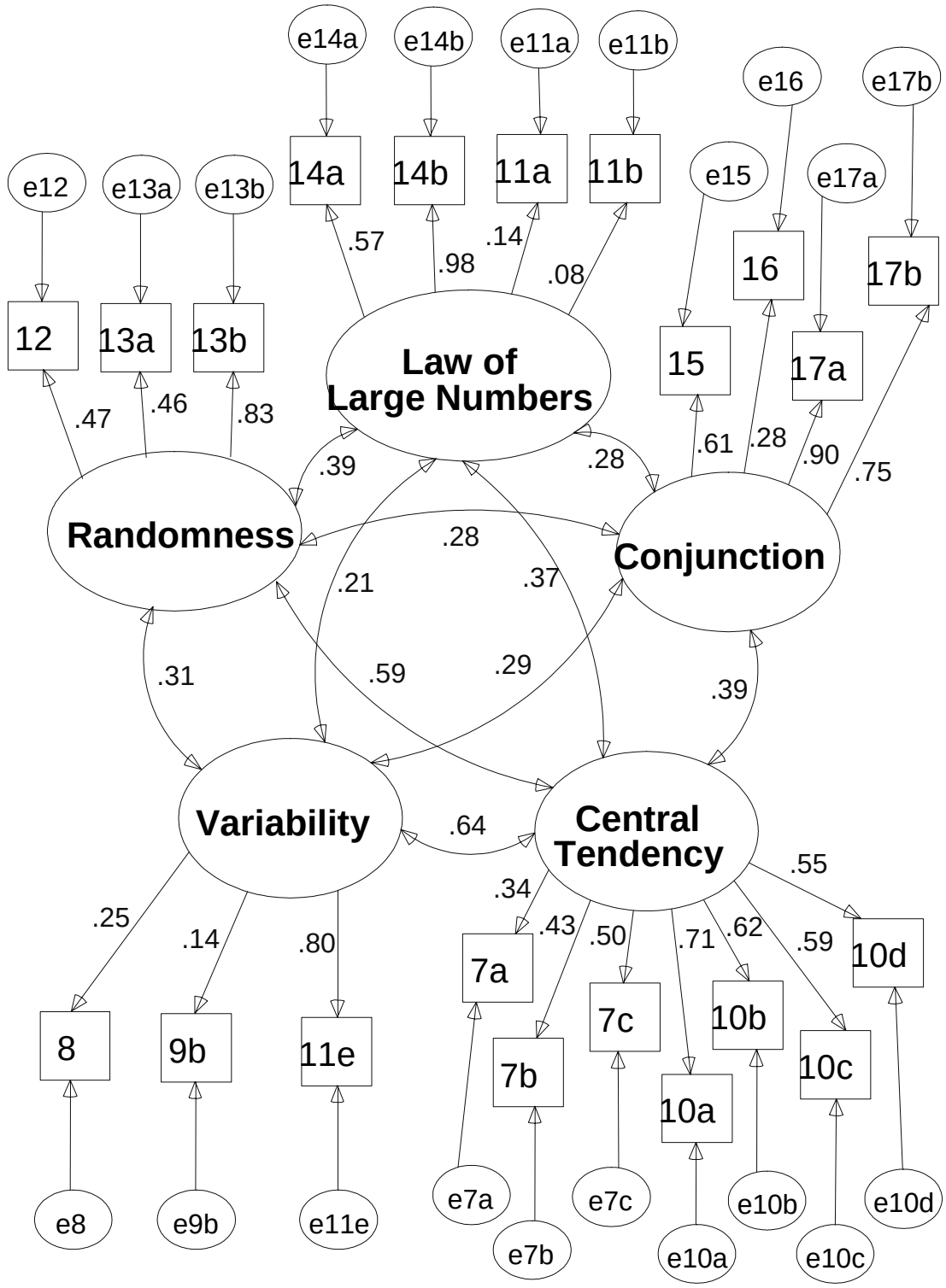


Figure 2. Standardized Regression Weights for Five-Factor Model.

Table 2
Path Coefficients for Measurement Model Relating Individual Items to Five Factors

Path	Unstandardized	Standard Error	Critical Ratio	P	Standardized
7a from Central Tendency (CT)	0.66	0.14	4.62	<.001	.34
7b from CT	0.60	0.10	5.89	<.001	.43
7c from CT	0.64	0.09	7.04	<.001	.50
8 from Variability	0.25	0.09	2.95	.003	.26
9b from Variability	0.23	0.14	1.67	.095	.14
10a from CT	0.74	0.07	10.63	<.001	.71
10b from CT	0.80	0.09	9.00	<.001	.62
10c from CT	0.84	0.10	8.49	<.001	.59
10d from CT	0.51	0.07	7.83	<.001	.55
11a from Law of Large Numbers (LLN)	0.21	0.11	1.92	.055	.14
11b from LLN	0.12	0.10	1.17	.242	.08
11e from Variability	0.91	0.19	4.88	<.001	.80
12 from Randomness	0.53	0.09	6.02	<.001	.47
13a from Randomness	0.38	0.07	5.90	<.001	.46
13b from Randomness	1.15	0.12	9.80	<.001	.83
14a from LLN	0.64	0.10	6.66	<.001	.58
14b from LLN	1.09	0.12	8.76	<.001	.98
15 from Conjunction	0.98	0.11	8.90	<.001	.61
16 from Conjunction	0.51	0.13	3.82	<.001	.28
17a from Conjunction	1.60	0.12	13.97	<.001	.90
17b from Conjunction	1.00	0.09	11.30	<.001	.75

Table 3
Covariances/Correlations for Measurement Model Relating Individual Items to Five Factors

	Covariance / Correlation	Standard Error	Critical Ratio	P
Randomness and Law of Large Numbers	.39	.08	4.60	<.001
Law of Large Numbers and Conjunction	.28	.08	3.63	<.001
Central Tendency and Conjunction	.39	.08	5.18	<.001
Central Tendency and Variability	.64	.13	4.77	<.001
Randomness and Variability	.31	.11	2.78	.005
Randomness and Central Tendency	.60	.08	7.86	<.001
Law of Large Numbers and Variability	.21	.10	2.26	.024
Law of Large Numbers and Central Tendency	.37	.08	4.62	<.001
Randomness and Conjunction	.28	.08	3.40	<.001
Variability and Conjunction	.29	.10	2.84	.005

The selected fit indices for this model were $\chi^2(179, N = 210) = 274.47$, $p < .001$, root mean square error of approximation (RMSEA) = .051, and comparative fit index (CFI) = .886. In search of paths that were not contributing to the model, the alpha-level was set at $\alpha = .05$. The standardized path coefficient for 11b was small (.083), and the path was not statistically significant ($p = .242$). Analysis of the results from this item indicated that less than 14% of the sample correctly answered this multiple-choice item, and over 63% chose the same incorrect answer. Therefore, item 11b was not contributing variance to the model and was removed from the composite score for the Law of Large Numbers and the total assessment composite score for further analyses but was explored individually where appropriate. The theoretical model was rerun without item 11b. The resulting model is displayed in Figure 3, the path

coefficients for the variables are listed in Table 4, and the correlations between latent variables are listed in Table 5.

Table 4
Coefficients for Measurement Model Relating Individual Items to Five Factors without 11b

Path	Unstandardized	Standard Error	Critical Ratio	P	Standardized
7a from Central Tendency (CT)	0.66	.14	4.62	<.001	.34
7b from CT	0.60	.10	5.89	<.001	.43
7c from CT	0.65	.09	7.04	<.001	.51
8 from Variability	0.25	.09	2.95	.003	.25
9b from Variability	0.23	.14	1.67	.095	.14
10a from CT	0.74	.07	10.64	<.001	.71
10b from CT	0.80	.09	8.99	<.001	.62
10c from CT	0.84	.10	8.49	<.001	.59
10d from CT	0.51	.07	7.84	<.001	.55
11a from Law of Large Numbers (LLN)	0.23	.11	2.06	.039	.15
11e from Variability	0.91	.19	4.87	<.001	.80
12 from Randomness	0.53	.09	6.02	<.001	.47
13a from Randomness	0.38	.07	5.90	<.001	.46
13b from Randomness	1.15	.12	9.77	<.001	.82
14a from LLN	0.66	.09	6.98	<.001	.60
14b from LLN	1.05	.12	8.98	<.001	.94
15 from Conjunction	0.98	.11	8.91	<.001	.61
16 from Conjunction	0.51	.13	3.83	<.001	.28
17a from Conjunction	1.60	.12	13.96	<.001	.90
17b from Conjunction	1.00	.09	11.30	<.001	.75

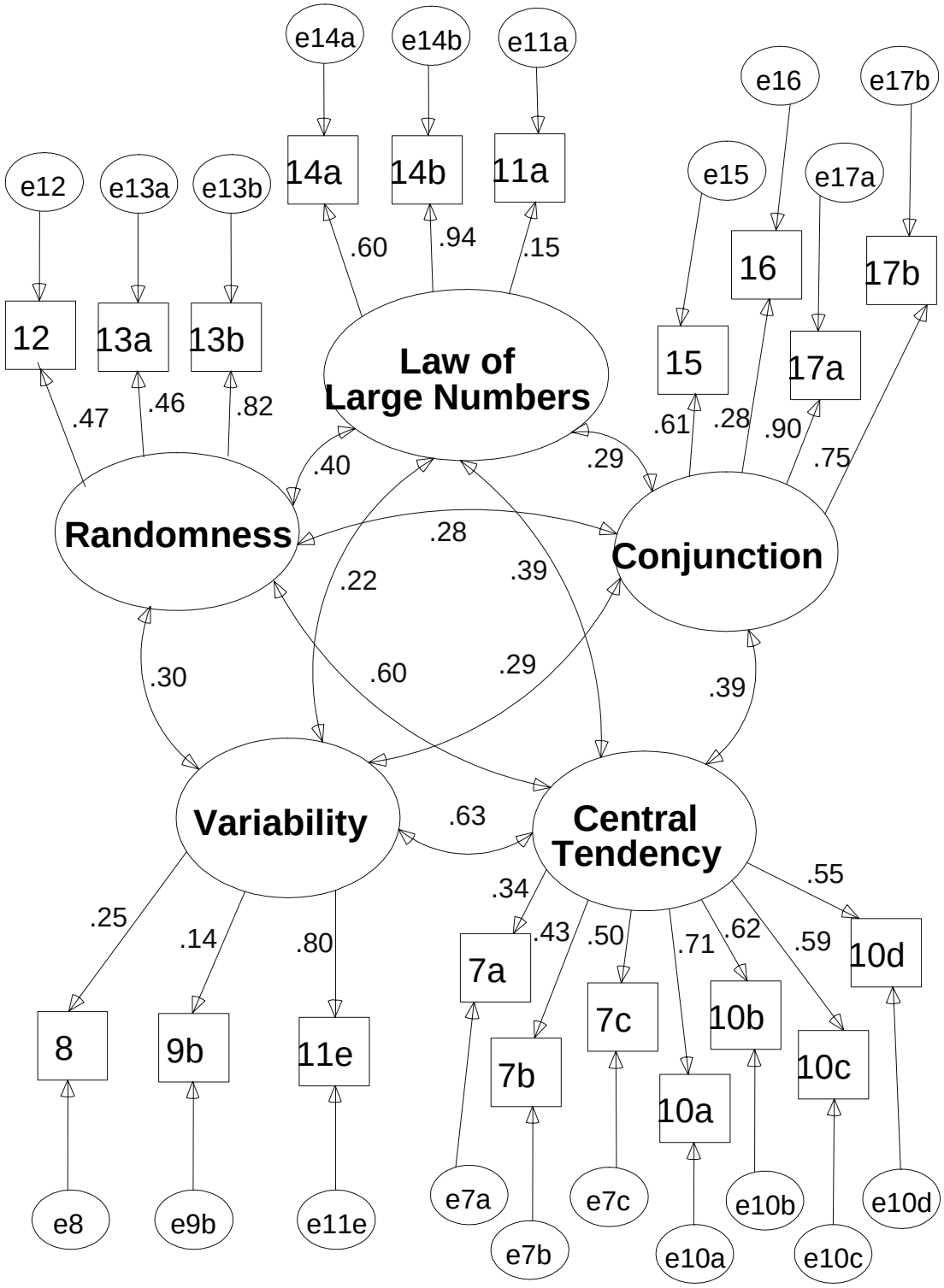


Figure 3. Standardized Regression Weights for Five-Factor Model without 11b.

Table 5
Covariances/Correlations for Measurement Model Relating Individual Items to Five Factors without 11b

	Covariance / Correlation	Standard Error	Critical Ratio	P
Randomness and Law of Large Numbers	.40	.09	4.69	<.001
Law of Large Numbers and Conjunction	.29	.08	3.78	<.001
Central Tendency and Conjunction	.39	.08	5.18	<.001
Central Tendency and Variability	.64	.13	4.76	<.001
Randomness and Variability	.31	.11	2.78	.005
Randomness and Central Tendency	.60	.08	7.87	<.001
Law of Large Numbers and Variability	.22	.10	2.29	.022
Law of Large Numbers and Central Tendency	.39	.08	4.82	<.001
Randomness and Conjunction	.28	.08	3.40	<.001
Variability and Conjunction	.29	.10	2.83	.005

The fit indices for this model were $\chi^2(160, N = 210) = 233.48$, $p < .001$, RMSEA = .047, and CFI = .910. The standardized path coefficient for 9b was fairly small (.139), and the path was not statistically significant ($\alpha = .05$, $p = .095$). Therefore, item 9b was analyzed for possible removal from the model. Item 9b assessed the concept of range, but the other two items comprising variability, 8 and 11e, assessed the concept of standard deviation. Removal of item 9b would remove the concept of range from the assessment diminishing the content validity of the assessment. Therefore, item 9b was retained in the model.

Based on this analysis of the measurement model, composite scores were formed for Randomness, Law of Large Numbers, Conjunction, Central Tendency, and Variability by summing the rubric scores for the associated items as described

previously with the exception of Law of Large Numbers which no longer utilized information from 11b, so it was composed of 11a, 14a, and 14b.

Utilizing the composite variables, the theoretical model was run with the data from the 210 PK-8 preservice teachers. Skewness and kurtosis values are listed in Table 6. The resulting model is displayed in Figure 4, and the path coefficients are displayed in Table 7. This model reveals that all five composite variables contributed significantly ($\alpha = .05$) to the model, and the two latent variables, Likelihood and Summary Data, were highly correlated. The fit indices for this model were $\chi^2(4, N = 210) = 4.46$, $p = .347$, RMSEA = .024, and CFI = .996.

Table 6
Assessment of Normality for Composite Variables and Time and Methods Model

	Skewness	Ratio of Skewness to Standard Error of Skewness	Kurtosis	Ratio of Kurtosis to Standard Error of Kurtosis
Randomness	-1.03	-6.16	0.94	2.81
Law of Large Numbers	0.03	0.20	0.56	1.66
Conjunction	0.55	3.26	-0.82	-2.44
Central Tendency	-0.99	-5.91	1.29	3.87
Variability	-1.13	-6.75	1.16	3.47

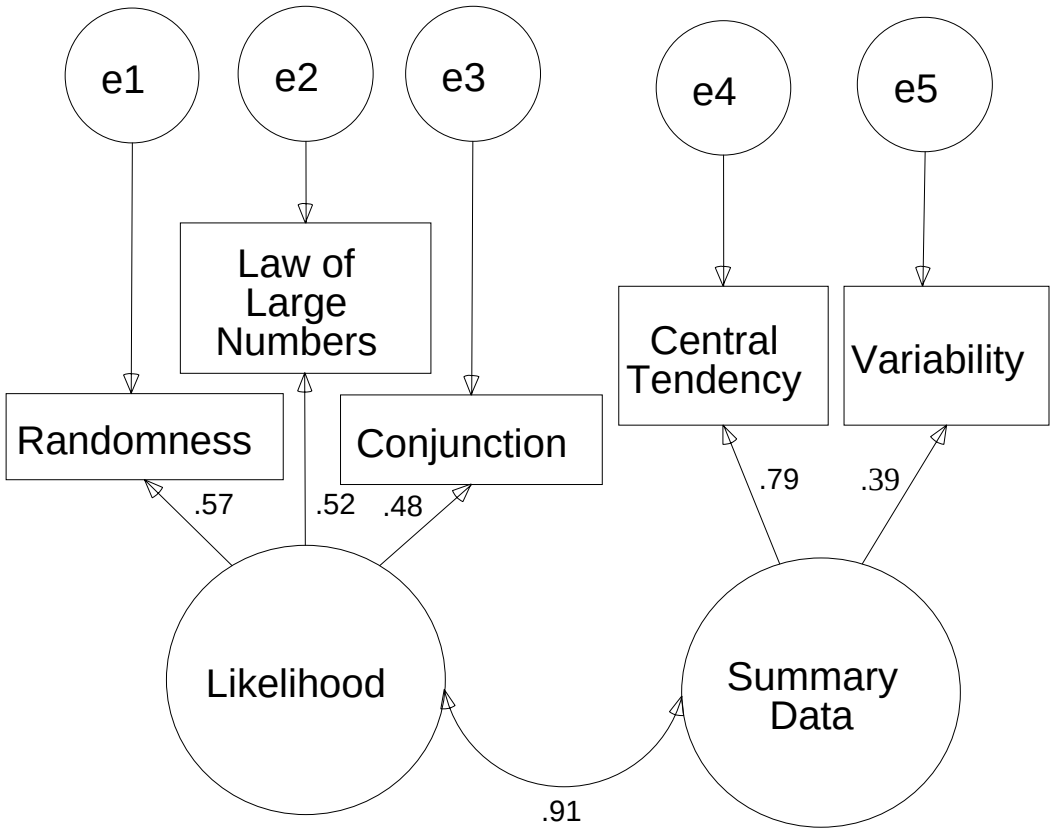


Figure 4. Theoretical Model with Standardized Coefficients.

Table 7
Theoretical Model Path Coefficients

Path	Unstandardized	Standard Error	Critical Ratio	P	Standardized
Random from Likelihood	1.44	0.20	7.13	<0.001	0.57
Law of Large Numbers from Likelihood	1.36	0.21	6.53	<0.001	0.52
Conjunction from Likelihood	2.32	0.38	6.03	<0.001	0.48
Central Tendency from Summary Data	4.57	0.62	7.33	<0.001	0.79
Variability from Summary Data	0.97	0.20	4.86	<0.001	0.39
Likelihood and Summary Data	0.91	0.12	7.65	<0.001	0.91

Analysis of Multiple-Choice and Short-Answer Responses

The question concerning the extent to which preservice PK-8 teachers were successful at answering multiple-choice and short-answer (non-extended-response) items concerning measures of center, measures of spread, and common misconceptions of chance was answered using descriptive statistics and frequencies. For consistency of data interpretation, all items were scored on an ordinal scale from zero to five. However, some multiple-choice and short-answer items did not have six levels of scoring. Therefore, item rubrics (see Appendix G) should be considered when the descriptive statistics are analyzed. Although the data are not strictly interval in nature, they approximate an interval scale. Therefore, statistics requiring interval scaling, such as mean and standard deviation, can be reasonably interpreted. Descriptive statistics including mean, median, mode, standard deviation, and quartiles are reported for each item in Table 8. Frequencies for each rubric score were computed for each item and are reported in Table 9.

Items for which less than a quarter of the participants answered correctly (11b, 14a, and 15) were investigated for commonalities. For all three of these items, the answer provided by the plurality of the participants was the choice indicating that options were equally likely. Further analysis of response frequencies revealed that equally likely was the most popular choice for every multiple-choice item containing that option (11a, 11b, 12, 13a, 14a, 15, and 17a), and 17 participants (8.1%) chose equally likely for all seven of these items. Further evidence of incorrect application of

the concept of equally likely can be seen in answers to extended-response items such as the following response to item 17b,

> You have no idea whether this woman is a doctor or a mother. Either she is or she isn't. So therefore the probability of her being a doctor is 0.50, and the probability of her being a mother is 0.50.

Table 8
Descriptive Statistics for Item Responses

				Standard	Percentiles		
	Mean	Median	Mode	Deviation	25	50	75
F7a Mean	3.61	5	5	1.91	1	5	5
F7b Median	4.37	5	5	1.41	5	5	5
F7c Mode	4.41	5	5	1.28	5	5	5
F8 Std MC	4.55	5	5	0.98	5	5	5
F9a Graph	4.34	5	5	1.44	5	5	5
F9b Range	3.87	5	5	1.67	3	5	5
F10a Mode	3.61	4	4	1.05	3	4	4
F10b Median	2.26	2	2	1.29	2	2	3
F10c Mean	2.40	2	2	1.42	2	2	4
F10d Typical	2.07	2	2	0.93	2	2	2
F10e Typical Type			5				
F11a Std Mean	4.27	5	5	1.52	5	5	5
F11b Std Dev	1.93	1	1	1.46	1	1	3
F11c Std Exp	2.61	3	3	1.17	2	3	3
F11d Std Dist	2.10	2	2	1.25	2	2	3
F11e STD	2.76	3	3	1.13	2	3	3
F12 Coin HT	4.60	5	5	1.13	5	5	5
F13a Birth	4.49	5	5	0.84	4	5	5
F13b Birth	2.97	3	2	1.39	2	3	4
F14a Coin Party	3.12	3	3	1.11	3	3	3
F14b Coin Party	2.33	2	2	1.11	2	2	2
F15 Dice	2.40	2	1	1.62	1	2	3
F16 Coin HH	3.21	4	5	1.82	1	4	5
F17a Doc	2.63	1	1	1.79	1	1	5
F17b Doc	2.55	2	2	1.34	2	2	3

Table 9
Frequencies for Non-Extended-Response Items

Score	Description	Frequency	Percent
7a Mean			
5	Mean is 85	137	65.2
1	Mean is NOT 85	73	34.8
7b Median			
5	81 is the Median	173	82.4
3	81 is the middle number provided	8	3.8
1	81 is NOT the median or middle number	29	13.8
7c Mode			
5	78 was the unique mode	166	89.0
4	78 was one of two modes	9	4.3
3	78 was one of five modes - all numbers were unique	11	5.2
2	78 was used but was not a mode	3	1.4
1	78 was not used	21	10.0
8 Std MC			
5	48 and 72	172	81.9
3	54 and 66	21	10.0
2	36 and 84	16	7.6
1	24 and 96	1	0.5
9a Graph			
5	Six	172	81.9
3	Four	7	3.3
1	Not 6 or 4	31	14.8
9b Range			
5	Eleven	138	65.7
3	Nine	25	11.9
1	A single number other than 9 or 11	47	22.4
11a Std Mean			
5	Approximately Equal	169	80.5
2	Less Than	19	9.0
1	More Than	14	6.7
0	Other Answers	8	3.8
11b Std Dev			
5	More Than	28	13.3
3	Less Than	44	21.0
1	Approximately Equal	133	63.3
0	Other Answers	5	2.4

Table 9 Continued

Score	Description	Frequency	Percent
12 Coin HT			
5	Equally Likely	186	88.6
2	Tails	12	5.7
1	Heads	12	5.7
13a Birth MC			
5	All Equally Likely	136	64.8
4	First 3 Equally Likely	52	24.8
3	GBGBGB	14	6.7
2	BGGBBG	5	2.4
2	GGGBBB	0	0.0
1	GGGGGB	3	1.4
14a Coin Party			
5	5 Flips - Kelly	39	18.6
3	Doesn't Matter - Shannon	145	69.0
1	5000 Flips - Casey	26	12.4
15 Dice			
5	5 on Red	50	23.8
3	5 on Red and Other than 6 on Blue	32	15.2
2	5 on Red and 6 on Blue	30	14.3
1	Doesn't Matter	98	46.7
16 Coin HH - MC			
5	0.16	84	40.0
4	0.36	33	15.7
3	0.25	12	5.7
2	0.8	19	9.0
1	0.6	8	3.7
1	0.4	41	19.5
0	1.2	4	1.9
0	0	1	0.5
0	1	1	0.5
0	0.5	7	3.3
17a Doc			
5	Doctor	68	32.4
3	Doctor and Mother	35	16.7
1	Equally Likely	107	51.0

Analysis of Extended Responses

The question concerning the quality of written explanations of PK-8 preservice teachers to items concerning measures of center, measures of spread, and common misconceptions of chance, was answered using descriptive statistics, frequencies, and representative participant responses from the data analysis based on content and pedagogical content knowledge literature. For consistency of data interpretation, all items were scored on an ordinal scale from zero to five. A zero represents no answer, a one represents a restatement of the question, a two represents an incorrect answer, a three represents an incomplete answer, a four represents an acceptable answer, and a five represents the desired answer. Although the data are not strictly interval in nature, they approximate an interval scale. Therefore, statistics requiring interval scaling, such as mean and standard deviation, can be reasonably interpreted. Descriptive statistics including mean, median, mode, standard deviation, and quartiles are reported for each item in Table 8 on page 39. Frequencies for each rubric score were computed for each item and were reported in Table 10.

Table 10
Frequencies for Extended-Response Items

Score	Description	Frequency	Percent
10a Mode			
5	Desired	20	9.5
4	Acceptable	132	62.9
3	Incomplete	33	15.7
2	Incorrect	16	7.6
1	Restate Question	0	0.0
0	Did Not Answer	9	4.3
10b Median			
5	Desired	7	3.3
4	Acceptable	45	21.4
3	Incomplete	3	1.4
2	Incorrect	125	59.5
1	Restate Question	1	0.5
0	Did Not Answer	29	13.8
10c Mean			
5	Desired	0	0.0
4	Acceptable	77	36.7
3	Incomplete	1	0.5
2	Incorrect	96	45.7
1	Restate Question	0	0.0
0	Did Not Answer	36	17.1
10d Typical			
5	Desired	1	0.5
4	Acceptable	23	11.0
3	Incomplete	4	1.9
2	Incorrect	163	77.6
1	Restate Question	0	0.0
0	Did Not Answer	19	9.0
10e Typical Type			
NA	Mode	126	60.0
NA	Median	11	5.2
NA	Mean	14	6.7
NA	Average - undefined	32	15.2
NA	Not a measure of Center	8	3.8
NA	Did Not Answer	19	9.0

Table 10 Continued

Score	Description	Frequency	Percent
11c Std Exp			
5	Desired	8	3.8
4	Acceptable	28	13.3
3	Incomplete	92	43.8
2	Incorrect	60	28.6
1	Restate Question	0	0.0
0	Did Not Answer	22	10.5
11d Std Dist			
5	Desired	5	2.4
4	Acceptable	12	5.7
3	Incomplete	67	31.9
2	Incorrect	83	39.5
1	Restate Question	1	0.5
0	Did Not Answer	42	20.0
11e STD			
5	Desired	2	1.0
4	Acceptable	49	23.3
3	Incomplete	96	45.7
2	Incorrect	43	20.5
1	Restate Question	0	0.0
0	Did Not Answer	20	9.5
13b Birth			
5	Desired	38	18.1
4	Acceptable	42	20.0
3	Incomplete	37	17.6
2	Incorrect	75	35.7
1	Restate Question	5	2.4
0	Did Not Answer	13	6.2
14b Coin Party			
5	Desired	23	11.0
4	Acceptable	9	4.3
3	Incomplete	4	1.9
2	Incorrect	160	76.2
1	Restate Question	7	3.3
0	Did Not Answer	7	3.3
17b Doc			
5	Desired	43	20.5
4	Acceptable	1	0.5
3	Incomplete	10	4.8
2	Incorrect	138	65.7
1	Restate Question	11	5.2
0	Did Not Answer	7	3.3

Responses to Randomness

Item 13b asked participants to explain their multiple-choice answer concerning the order in which six children are most likely to be born. Thirty-eight of the participants (18.1%) provided a desired answer by explaining that each birth was independent of the previous births (e.g., “For each child, there is a 50% chance of having a boy or girl. The sex of the previous child has no effect on the next, so all of the choices are equally likely” or

> There are only two possibilities, either a boy will be born or girl will be born. Each possibility is independent of the other. Each time a child is born there is a 50% chance it will be a girl and a 50% it will be a boy and the previous child born does not effect the sex of the current child.

One senior preservice teacher pursing certification in grades 4-8 with a specialty in mathematics and science provided extraordinary detail with the following response:

> If order matters, then they are all equally probable. Let's simplify the problem. If we had only one child, is a boy or girl more likely? They are equally probable. If we have two children, are you more likely to have a boy and then a girl, or a girl and then a boy, a boy and another boy, or a girl and another girl? Again, they are equally probable. In fractions, you have a 1/2 chance for the first (no matter which gender) and a 1/2 chance for the second (no matter which gender). When you multiply this out, you get 1/4. Each of the four choices has a 1/4 chance of happening (which makes sense, because when you add it you get 1 and we've covered all the options). If we carry this out all the way to 6 children, the same

> principle is true. You have a 1/2 chance for each child (no matter if we're talking about a boy or a girl) and multiplied out for each child you get 1/64, for each of the 64 choices (only four of which are listed).

Forty-two additional participants (20%) provided an acceptable explanation by explaining that there was a 50% chance for each birth of the child being the chosen gender. These participants did not specifically mention that previous births have no effect on future births (e.g., “Each child has a 50% chance of being a boy and a 50% chance of being a girl. All choices are equally likely”). Thirty-seven participants (17.6%) provided incomplete answers. Most of these students referred to probabilities for a single birth, and did not make a connection to the sequence of births (e.g., “There is a 50/50 chance for a new born baby to be a boy or a girl”). Seventy-five participants (35.7%) provided incorrect explanations. Most of these (32 of the 75) interpreted the equal probability of each gender for a particular birth to mean that there should be an equal number of boys and girls in the family, for example:

> Since each child has equal probability of being a boy or a girl (0.5), then the total probability is 0.5 of being a boy or a girl. The first three choices show a 0.5 ratio of boys to girls. The last choice shows a 0.83 chance of being a girl and a 0.17 chance of being a boy.

and

> The chances for having a boy and girl are 50 - 50, the order isn't important. So the first three choices are all equally likely and more like than the fourth because the fourth isn't representing the 50 - 50 chances of having a boy or a girl.

Some even took the concept of equal probability a step further and expected the births to maintain an equal number of each gender throughout the birth process:

> There is a fifty/fifty chance of either a girl or a boy being born at each instance. Therefore the order must be girl-boy-girl-boy to show that there is an equal chance of either a boy or a girl being born.

Sixteen of the 75 participants who provided incorrect answers abandoned statistical reasoning and based their answers on the context of the situation (e.g., “The chromosome in the male is going to determine the sex of the children in the family” and “How many children you have and the sex of the children is all in God's hands”). Five participants (2.4%) simply repeated their selection to the multiple-choice part of the question as their explanation (e.g., “the probability is the same for all choices”), and 13 participants (6.2%) either did not answer the question or said that they did not know the answer.

Responses to Law of Large Numbers

Item 14b asked participants to explain their reasoning for how many times the principal should flip the coin to give the class the best chance of getting tails at least 60% of the time. Though none of the participants actually mentioned the law of large numbers by name, 36 (17.1%) of the participants correctly utilized the law of large number reasoning at some level. Twenty-three (11%) of the participants provided a desired explanation by explaining that the actual results of the flipping experiment would approach the expected theoretical probability of 50% tails as the number of coin flips increases, so the class would be more likely to get a result of 60% tails if they chose

the smaller number of flips (e.g., “its easier to get 60 percent with a smaller number, because with more flips, it will be closer to 50 percent”). Nine of the participants (4.3%) provided acceptable answers based on the law of large number reasoning, but they did not carry the reader through the argument of which choice to make (e.g., “the more flips you make the closer you'll get to the mean, or the 50/50 that a coin flip would produce”). Four (1.9%) of the participants gave incomplete answers that hinted at the use of the law of large numbers but were ambiguous (e.g., “The coin would need to land on tails 3 out of the 5 times. The more chances you give it the less likely it will land on tails 60% of the time”). The vast majority of the participants, 160 (76.2%), provided incorrect explanations. Most of these, 92 (43.8% of the total sample) simply referred to the 50% chance of flipping a tail but did not explain how this related to the overall probability of the flipping experiment (e.g., “There is always a 50% chance of either side of the coin”). Twenty of the incorrect responses (9.5% of the total sample) referred to equal ratios for both cases, for example,

> In order to get 60% tails according to Casey's 5000 times, you would have to get tails 3000 out of 5000 as tails. This fraction reduced is 3/5. Kelly suggests 5 times which means that 3 out of the 5 times would have to be tails. Therefore, by saying that they have the same chance, Shannon is correct.

Seven participants (3.3%) restated their answer choice as their explanation, and seven participants (3.3%) either did not answer or said that they did not know the answer.

Responses to Conjunction

Item 17b asked participants to explain their reasoning regarding the description of the woman holding the baby. Forty-three (20.5%) of the participants provided a desired answer by explaining that being a doctor and a mother is less likely than being a doctor because the group composed of women who are a doctor and a mother is a subset of the group of women who are a doctor (e.g., "The chances of someone meeting 1 specific criteria, that of being a doctor, is more likely than meeting both a doctor and a mother"). One of these participants used an example from science to illustrate this concept:

> It is more likely to have just one thing than a specific combination of both. Because then you are not just finding out how likely one thing, or how likely the other thing is, but BOTH at the SAME TIME. This to me is like recessive and dominant traits. For Bb and Bb, the likelihood that a child receives B is at least 50%, or b is 50%, but the likelihood that a child receives both BB is 25%.

One participant (0.5%) provided an acceptable explanation by explaining the computations that would be involved in computing the probability of being a doctor and a mother if the individual probabilities of being a doctor and being a mother were known:

> For the woman to be a doctor and a mother you multiply the probability of her being a doctor by the probability of her being a mother to get the likelihood of her being both. Since both numbers are a fraction, you get a smaller when you

> multiply two fractions, thus the probability of her being a doctor alone is higher than her being both.

Ten participants (4.8%) provided answers using the same reasoning processes as previously described answers, but were vague in their explanations (e.g. “I would say the first one is more likely because it is involving only one variable”).

The majority (138, which is 65.7%) of the participants provided incorrect explanations. The vast majority of these (126, which is 60.0% of the entire sample) provided context-driven explanations such as “It is more likely that she is both because of the baby. Most of the time if someone is carrying a baby it means that the baby is theirs” and

> Because we do not know the woman and there is no information on her clothes or the location of where she is carrying the baby to indicate her profession, we have only to assume it is equally likely to be a doctor or a doctor and a mother.

Eleven participants (5.2%) restated their answer choice as their explanation, and seven participants (3.3%) either did not answer or said that they did not know the answer.

Responses to Central Tendency

Item 10 presented respondents with a bar graph and asked them to explain how the word *typical* was interpreted by three sample students based only on the numeric response of these students. Then item 10 asked the participants which measure of typicality they would use. The three sample student responses represented calculations of mode, median, and mean, respectively. Forty-three (20.5%) of the participants were able to correctly identify the computational method for all three measures of typicality.

Forty-five (21.4%) participants identified two of the three measures; 99 (47.1%) correctly identified only one of the measures, and 23 (11%) did not identify any of the measures of typicality.

Most of the participants (185, which is 88.1%) were able to correctly (rubric score of 3 or higher) determine that the sample student who used zero as the typical number of siblings had used mode as the measure of typicality (see Table 10 on page 43 for a finer disaggregation of the data). Fifty-three (25.2%) of the participants were able to correctly identify a response of *one* as the measure of typicality represented by use of the median and two more (1.0%) correctly said that this answer could be achieved by rounding the mean after removing the outliers; however, 35 (16.7%) of the participants used modal concepts, 25 (11.9%) used the word *average* without designating which definition of average they intended to use, 16 (7.6%) used the concept of mean without mention of removing the outliers, and 20 (9.5%) used non-statistical answers to interpret the sample student's response (e.g., "Maybe Jennifer considers herself 'typical' and only has 1 sibling herself"). Seventy-eight (37.1%) of the participants correctly identified the sample student response of *two* as the typical number of siblings represented by use of the concept of mean. Of these 78 participants, 28 used the word *average* without indicating which definition of average was intended.

In item 10d, the participants were asked to explain which measure of typicality they would use for this data. Quality of explanations were coded in 10d and choice of typicality was recorded in 10e. The majority of the participants (126, which is 60.0%) chose mode, 11 (5.2%) chose median, 14 (6.7%) chose mean, 32 (15.2%) chose

"average" without explaining which meaning of average they were using. Eight participants (3.8%) did not use a measure of typicality, and 19 (9.0%) did not answer the question or said that they did not know. However, only 28 participants (13.3%) were able to provide a reason for their choice other than defining their choice. Most (22) of these reasons were based on where the majority of the data were located or on the outliers. Sixty-six (31.4%) of the participants answered the part of the item telling which measure they would use but did not explain why. Of those who tried to explain their reason, 66 (31.1% of the entire sample) only stated their choice in multiple ways or defined their choice rather than actually explaining reasons beyond those requested in the previous parts of item 10 (e.g., "I would use 0 siblings because that was the most common recording").

Responses to Variability

Item 11c was intended to elicit a conceptual definition for standard deviation and item 11d was intended to elicit information about the percentage of data encompassed within one, two, or three standard deviations of the mean under a normal curve. However, participant responses did not align with the expected outcomes for the items (i.e., some participants answered item 11c in 11d or in both 11c and 11d or vice-versa). Therefore, for scoring purposes items 11c and 11d were scored together and reported as 11e. On item 11c, 11d, or both, 75 (35.7%) of the participants mentioned only variability when describing standard deviation (e.g. "it is how spread out the data is from the mean"). Fifty of the participants (23.8%) only described how the scores would be distributed around the mean for a normal distribution (e.g., "Approximately 68% of the

360 total scores should fall between a 64 and an 86"). Twenty-one (10.0%) of the participants described both variability and the distribution of scores. Eight of these participants who mentioned both variability and the distribution (3.8% of the entire sample) specifically mentioned that the standard deviation describes how the data varies around the mean (e.g. "Its an average of how far off people were from the mean score", as opposed to those who only defined deviation or those who did not mention the mean as the center of the variation). Forty-four (21.0%) of the participants who provided an explanation did not correctly describe standard deviation. Twenty additional participants (9.5%) did not attempt to explain standard deviation.

Observations across Extended Responses

Six of the participants scored at least a three (incomplete, acceptable, or desired answers) on the extended-response items for all five measured variables (randomness, law of large numbers, conjunction, central tendency (all three parts), and variability). Eighteen participants scored at least a three on four of the measured variables, 30 participants scored this well on three of the measured variables, 78 for two of the measured variables, 49 for one of the measured variables, 18 scored at least a three on parts of the central tendency item, and 11 did not score this well on any of the extended-response items.

Comparison of Multiple-Choice and Short-Answer Responses to Extended Responses

Extended-response items were matched with the multiple-choice or short-answer items testing similar content. Seven extended-response items (10a, 10b, 10c, 11e, 13b, 14b, and 17b) requested an explanation of concepts tested using multiple-choice or

short-answer items (7c, 7b, 7a, 8, 13a, 14a, and 17a, respectively). Because all items were scored on rubrics with the same scale and the number of items in each of these comparison groups was equal, a composite score for each participant was formed for extended-response items by summing the rubric scores for 10a, 10b, 10c, 11e, 13b, 14b, and 17b. A composite score for each participant was similarly formed for non-extended-response (multiple-choice and short-answer) items by summing the rubric scores for 7c, 7b, 7a, 8, 13a, 14a, and 17a. The data are approximately interval in nature and have similar variances (see Table 11). Although the multiple-choice items are negatively skewed and extended-response items have a more peaked shape than normal, the two variables are not skewed in opposite directions, so they do not cause great concern (Sheskin, 2004). The Pearson's *r* correlation between these composite variables was calculated to be $r = .56$ with $p < .001$ (the non-parametric counterpart, a Spearman's *rho* correlation yielded $r_s = .52$, with $p < .001$) for $N = 210$. A dependent samples *t*-test indicated a statistically significant ($\alpha = .05$) difference between the scores for participants on the extended-response ($M = 18.89$, $SD = 5.31$) and non-extended-response ($M = 30.63$, $SD = 4.92$) portions of these items with higher scores on the non-extended-response portions ($t(209) = 35.27$, $p < .001$, $d = 2.29$). The non-parametric counterpart to the *t*-test for two dependent samples, a Wilcoxon Matched-Pair Signed-ranks test, indicated the same conclusion ($z = -12.54$, $p < .001$).

Table 11
Descriptive Statistics for Composite Variables for Similarly Matched Items

	Mean	Standard Deviation	Skewness	Ratio of Skewness to Standard Error of Skewness	Kurtosis	Ratio of Kurtosis to Standard Error of Kurtosis
MC Similar	30.63	4.92	-0.82	-4.90	0.36	1.08
ER Similar	18.89	5.31	-0.19	-1.14	0.92	2.74

Note. Total possible composite score is 35.

For a subset of these items, the content match was exact because the extended-response items (13b, 14b, and 17b) requested an explanation of the participants' responses to the multiple-choice portion (13a, 14a, and 17a, respectively). The same calculations were performed using composite scores formed from only these items. The data are approximately interval in nature and have similar variances (see Table 12). Although the extended-response items are positively skewed and have a more peaked shape than normal, the two variables are not skewed in opposite directions, so they do not cause great concern (Sheskin, 2004). The Pearson's *r* correlation between these composite variables was calculated to be $r = .65$ with $p < .001$ (the non-parametric counterpart, a Spearman's *rho* correlation yielded $r_s = .62$, with $p < .001$) for $N = 210$. A dependent samples *t*-test indicated a statistically significant ($\alpha = .05$) difference between the scores for participants on the extended-response ($M = 7.86$, $SD = 2.68$) and multiple-choice ($M = 10.24$, $SD = 2.49$) portions of these items with higher scores on the multiple-choice portions, $t(209) = 15.87$, $p < .001$, $d = 0.92$. The non-

parametric counterpart to the *t*-test for two dependent samples, a Wilcoxon Matched-Pair Signed-ranks test, indicated the same conclusion ($z = -11.12$, $p < .001$).

Table 12
Descriptive Statistics for Composite Variable for Exactly Matched Items

	Mean	Standard Deviation	Skewness	Ratio of Skewness to Standard Error of Skewness	Kurtosis	Ratio of Kurtosis to Standard Error of Kurtosis
MC Exact	10.24	2.49	0.02	0.10	-0.45	-1.35
ER Exact	7.86	2.68	0.49	2.94	1.06	3.16

Note. Total possible composite score is 15.

For each of the items that were matched exactly, the extended-response answers of the participants who correctly answered the multiple-choice were explored and are reported in Table 13.

Table 13
Comparisons of Non-Extended-Response and Extended-Response Answers

	Item 13	Item 14	Item 17
Number who correctly answered multiple-choice	136	39	68
Percent of $N = 210$ who correctly answered multiple-choice	64.8%	18.6%	32.4%
Number who correctly answered multiple-choice and provided acceptable or desired explanation	74	29	44
Percent of $N = 210$ who correctly answered multiple-choice and provided acceptable or desired explanation	35.2%	13.8%	21.0%
Percent of those who correctly answered multiple-choice who also provided acceptable or desired explanation	54.4%	74.4%	64.7%

Comparison of Responses Based on Certification Level Pursued

The participants in this study were all pursuing certification to teach, but had chosen one of three tracks: grades EC-4, grades 4-8 specializing in Language Arts and Social Studies, or grades 4-8 specializing in Mathematics and Science. Two comparisons were of interest: comparing performance on the two latent factors in the theoretical model, Likelihood and Summary Data, and comparing performance on the two types of items utilized in the assessment, extended-response and non-extended-response (short-answer and multiple-choice).

Latent Factors – Likelihood and Summary Data

The theoretical model for the assessment used in this study contained two latent factors, Likelihood and Summary Data, representing the two primary foci of the NCTM and INTASC standards, probability and data analysis, respectively. Therefore, composite scores were formed for each of these variables by summing the rubric scores for the items corresponding to the measured variables associated with the latent variables (i.e., 11a, 12, 13a, 13b, 14a, 14b, 15, 16, 17a, and 17b corresponding to Likelihood and 7a, 7b, 7c, 8, 9b, 10a, 10b, 10c, 10d, 11e corresponding to Summary Data, notice that 11b was removed from the model during the original model analysis) resulting in a possible composite score of 50 for each latent variable. To investigate potential differences in performance on the assessment among these three groups, a one-way analysis of variance (ANOVA) was run with certification route as the independent variable. The ANOVA did indicate a statistically significant ($\alpha = .025$ using Bonferroni corrections; Huck, 2004) difference among the three certification plans on the composite variable for

likelihood, $F(2, 207) = 8.17$, $p < .001$, and $\eta_p^2 = .073$, and the composite variable for summary data, $F(2, 207) = 5.86$, $p = .003$, and $\eta_p^2 = .054$. A Kruskal-Wallis ANOVA by ranks (the non-parametric counterpart) yielded similar results – a statistically significant difference among the three certification plans on the composite variable likelihood, $\chi^2(2, N = 210) = 16.19$, $p < .001$, and the composite variable for summary data, $\chi^2(2, N = 210) = 10.48$, $p = .005$.

To investigate the nature of these differences, two contrasts were formed. The first contrast was based on the idea that those preparing for the same grade bands would perform similarly and grouped those seeking certification to teach 4-8 Language Arts and Social Studies with those seeking certification to teach 4-8 Mathematics and Science and compared them to the participants preparing to teach EC-4. A *t*-test for two independent samples did not indicate a statistically significant ($\alpha = .0125$ using Bonferroni corrections; Huck, 2004) difference between those seeking certification for EC-4 ($M = 31.69$, $SD = 7.42$) and those seeking certification for grades 4-8 ($M = 33.57$, $SD = 7.19$) on the composite variable for likelihood, $t(208) = -1.86$, $p = .064$, $d = .26$, or the composite variable for summary data, (EC-4 $M = 32.98, SD = 7.71$; 4-8 $M = 34.94$, $SD = 5.96$), $t(202.9) = -2.07$ with equal variances not assumed ($p = .040$, $d = .28$). The Mann-Whitney U test for two independent samples (a non-parametric counterpart to the *t*-test for two independent samples which yields identical results to a Kruskal Wallis one-way analysis of variance on two samples) yielded similar results (no statistically significant difference between

those seeking certification for EC-4 and those seeking certification for grades 4-8 on the composite variable for likelihood, $z = -1.53$, $p = .126$, or the composite variable for summary data, $z = -1.62$, $p = .106$).

The second contrast was based on the idea that those specifically preparing to teach mathematics and science would outperform others on a test of probability and statistics topics and grouped those seeking certification to teach 4-8 Language Arts and Social Studies with those seeking certification to teach EC-4 and compared them to the participants preparing to teach 4-8 Mathematics and Science. A *t*-test for two independent samples did indicate a statistically significant ($\alpha = .0125$ using Bonferroni corrections; Huck, 2004) difference between those seeking certification for 4-8 Mathematics and Science ($M = 36.61$, $SD = 6.50$) and those seeking certification for one of the other two areas ($M = 31.61$, $SD = 7.23$) on the composite variable for likelihood, $t(208) = 4.05$, $p < .001$, $d = .73$, and the composite variable for summary data (Mathematics and Science $M = 37.17$, $SD = 5.35$; Others $M = 33.12$, $SD = 7.12$), $t(208) = 3.41$, $p = .001$, $d = .64$. Similar results were indicated by the Mann-Whitney U test for two independent samples (a statistically significant difference between those seeking certification for 4-8 Mathematics and Science and those seeking certification for one of the other two areas on the composite variable for likelihood, $z = -3.98$, $p < .001$ and the composite variable for summary data, $z = -3.24$, $p = 0.001$.

Item Type – Extended-Response and Non-Extended-Response

The assessment was formed using two primary item types, extended-response and non-extended-response (i.e., multiple-choice and short-answer). Composite scores

were formed for each of these item types by summing the rubric scores for the items corresponding to each type (i.e., 10a, 10b, 10c, 10d, 11e, 13b, 14b, and 17b for extended-response and 7a, 7b, 7c, 8, 9b, 11a, 12, 13a, 14a, 15, 16, and 17a for non-extended-response) resulting in a possible composite score of 40 for extended-response and 60 for non-extended-response. To investigate potential differences in performance on the assessment among these three certification groups, a one-way analysis of variance (ANOVA) was run with certification route as the independent variable. The ANOVA did indicate a statistically significant ($\alpha = .025$ using Bonferroni corrections; Huck, 2004) difference among the three certification plans on the composite variable for extended-response, $F(2, 207) = 6.65$, $p = .002$, and $\eta_p^2 = .060$, and the composite variable for non-extended-response, $F(2, 207) = 8.36$, $p < .001$, and $\eta_p^2 = .075$. A Kruskal-Wallis ANOVA by ranks (the non-parametric counterpart) yielded similar results – a statistically significant ($\alpha = .05$) difference among the three certification plans on the composite variable for extended-response, $\chi^2(2, N = 210) = 9.70$, $p = .008$ and the composite variable for non-extended-response $\chi^2(2, N = 210) = 16.16$, $p < .001$.

To investigate the nature of these differences, the two contrasts used for the previous analysis were used again. A *t*-test for two independent samples did not indicate a statistically significant, $\alpha = .0125$ using Bonferroni corrections (Huck, 2004), difference between those seeking certification for EC-4 ($M = 20.18$, $SD = 5.92$) and those seeking certification for grades 4-8 ($M = 21.82$, $SD = 5.52$) on the composite variable for extended-response, $t(208) = -2.07$, $p = .04$, $d = .29$, or the composite

variable for non-extended-response (EC-4 $M = 44.49, SD = 8.53$; 4-8 $M = 46.69$, $SD = 7.18$), $t(208) = -2.01$, $p = .05$, $d = .05$. The Mann-Whitney U results were similar (no statistically significant difference between those seeking certification for EC-4 and those seeking certification for grades 4-8 on the composite variable for extended-response, $z = -1.49$, $p = .136$, or the composite variable for non-extended-response, $z = -1.632$, $p = 0.103$. A *t*-test for two independent samples did indicate a statistically significant, ($\alpha = .0125$ using Bonferroni corrections (Huck, 2004), difference between those seeking certification for 4-8 Mathematics and Science ($M = 23.83$, $SD = 5.48$) and those seeking certification for one of the other two areas ($M = 20.27$, $SD = 5.65$) on the composite variable for extended-response, $t(208) = 3.64$, $p < .001$, $d = .64$, and the composite variable for non-extended-response (EC-4 $M = 49.95, SD = 5.60$; 4-8 $M = 44.47$, $SD = 8.10$), $t(85.7) = 5.11$ with equal variances not assumed, $p < .001$, $d = .79$. The non-parametric results were similar. The Mann-Whitney U test for two independent samples did indicate a statistically significant difference between those seeking certification for 4-8 Mathematics and Science and those seeking certification for one of the other two categories on the composite variable for extended-response ($z = -3.11$, $p = .002$) and the composite variable for non-extended-response ($z = -3.99$, $p < .001$).

Effects of Cumulative Exposure

The question concerning the effect of cumulative course exposure to probability and statistics topics on PK-8 preservice teachers' ability to determine a correct answer

and their ability to explain that answer for items concerning likelihood and summary data, structural equation models were explored.

Because participants would be expected to perform better on a probability and statistics assessment when the lag time between the assessment and an elementary statistics course was small, participants were awarded a 3 for the variable time since elementary statistics if they were taking elementary statistics during the semester of the assessment (44.3% of the sample), a 2 if they took elementary statistics in the year (Fall, Spring, or Summer) prior to the assessment (35.7%), and a 1 if they took elementary statistics more than a year before the assessment (20.0%). The three methods courses, ECFB 440, MEFB 450, and MEFB 460 were counted toward this total primarily for their influence on the pedagogical knowledge of the participants. For simplicity in terms, content courses taught through the education department such as MASC 351 and MASC 450 were counted toward the number of methods courses for this analysis because they offer an opportunity to revisit probability and statistics topics from an educational perspective. Of the 210 participants, 45.7% had not taken any of the specified methods courses yet, 44.3% had taken one, 0.5% had taken two, 6.2% had taken three, and 3.3% had taken four. Values for number of methods courses and time since elementary statistics for skewness (1.72 and -0.44 with ratios to standard error of skewness of 10.25 and -2.62, respectively) and kurtosis (2.808 and -1.170 with ratios to standard error of kurtosis of 8.40 and -3.50, respectively) indicate that number of methods courses is positively skewed and peaked whereas time since elementary statistics is negatively skewed. The model with standardized path coefficients is presented in Figure 5.The fit

indices for this model were $\chi^2(11, N = 210) = 128.72$, $p < .001$, RMSEA = 0.226, CFI = 0.556. These fit indices indicate that the data do not fit this model well. All of the path coefficients were statistically significant at $\alpha = .05$ and of a reasonable size (see Table 14); therefore, no paths were removed. The only change suggested by the modification indices was to correlate time since elementary statistics and number of methods courses, but this change caused a negative covariance estimate indicating that the model was not appropriate.

Table 14
Time and Methods Model Path Coefficients

Path	Unstandardized	Standard Error	Critical Ratio	P	Standardized
Likelihood from Time Since Statistics	0.75	.30	2.54	0.011	.23
Summary Data from Number of Methods Courses	1.78	.39	4.58	<0.001	.35
Likelihood from Number of Methods Courses	1.11	.26	4.34	<0.001	.43
Summary Data from Time Since Statistics	2.42	.50	4.86	<0.001	.37
Random from Likelihood	0.57	.11	5.46	<0.001	.57
Conjunction from Likelihood	1.75	.32	5.46	<0.001	.52
Central Tendency from Summary Data	4.62	.94	4.90	<0.001	.81
Variability from Summary Data	0.22	.04	4.90	<0.001	.43
Law of Large Numbers from Likelihood	0.61	.11	5.52	<0.001	.58

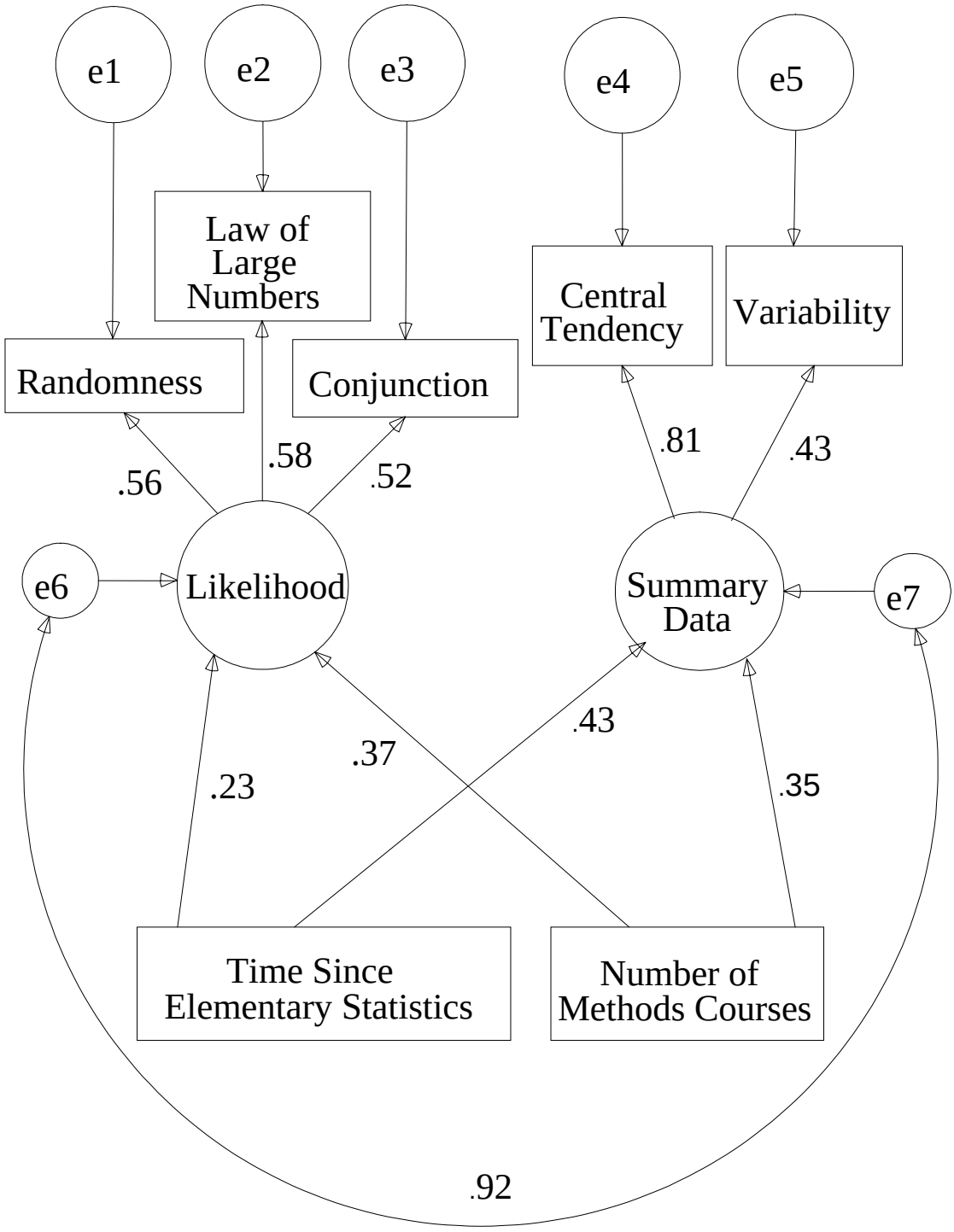


Figure 5. Time and Methods Model with Standardized Regression Coefficients.

Because the correlation between Likelihood and Summary Data was large in the theoretical model, a single-factor model was tested in preparation for an alternative time and methods model. The standardized regression weights for the model are displayed in Figure 6, and the path coefficients are displayed in Table 15. No modifications were suggested for this model by modification indices, and all the paths were significant at $\alpha = .05$ and of reasonable size. The fit indices for this model were $\chi^2(5, N = 210) = 5.025$, $p = 0.413$, RMSEA = .005, CFI = 1.000. Notice that this model and the theoretical model have similar fit indices in ranges that indicate that the data fit the models well.

The model in Figure 7 depicts addition of the variables time since elementary statistics and number of methods courses to the model with standardized path coefficients. The fit indices for this model were $\chi^2(14, N = 210) = 135.41$, $p < .001$, RMSEA = .204, and CFI = .543. These fit indices indicate that the data do not fit this model well. All of the path coefficients were statistically significant at $\alpha = .05$ and of a reasonable size (see Table 16); therefore, no paths were removed. The primary change suggested by the modification indices was to correlate time since elementary statistics and number of methods courses, but this change caused a negative covariance estimate indicating that the model was not appropriate. Therefore, other statistical methods were used to further investigate this research question.

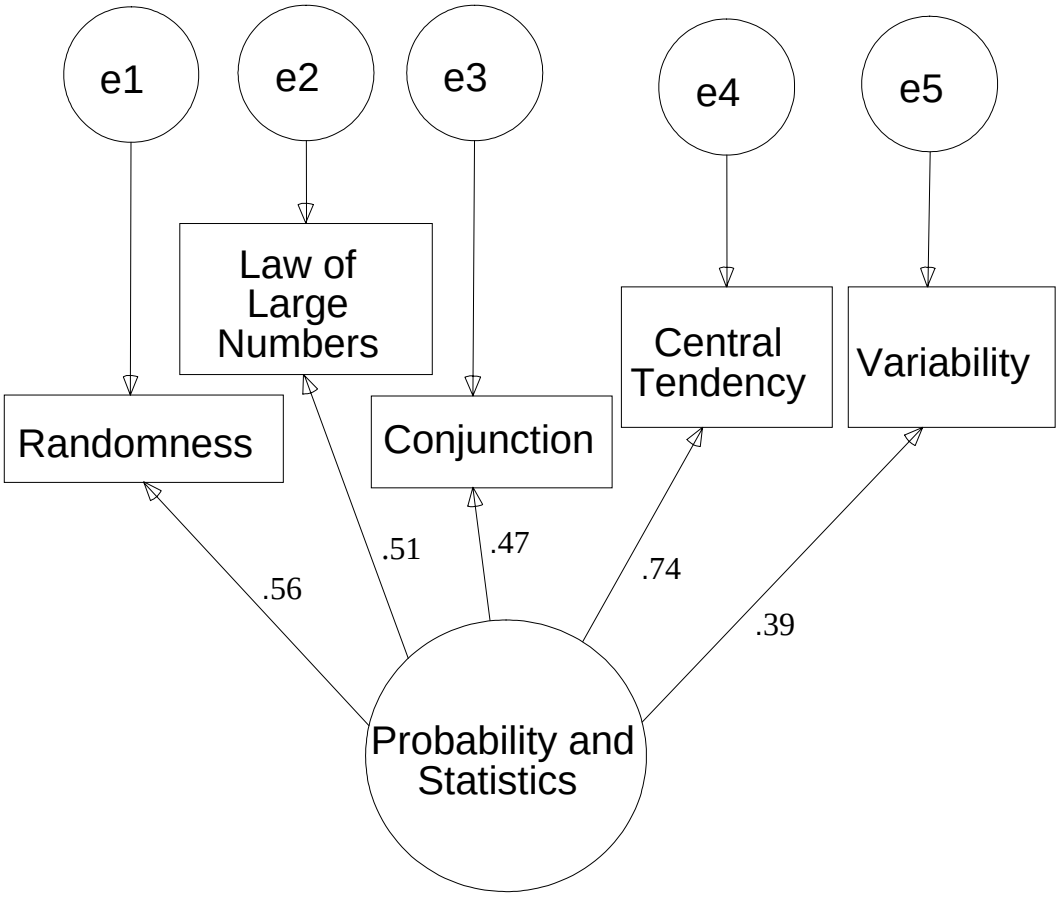


Figure 6. Single-Factor Model with Standardized Regression Coefficients.

Table 15
Single-Factor Model Path Coefficients

Path	Unstandardized	Standard Error	Critical Ratio	P	Standardized
Randomness from Probability and Statistics	1.41	0.20	7.18	<0.001	0.56
Conjunction from Probability and Statistics	2.28	0.38	6.03	<0.001	0.47
Central Tendency from Probability and Statistics	4.28	0.46	9.31	<0.001	0.74
Variability from Probability and Statistics	0.97	0.20	4.91	<0.001	0.39
Law of Large Numbers from Probability and Statistics	1.34	0.21	6.52	<0.001	0.51

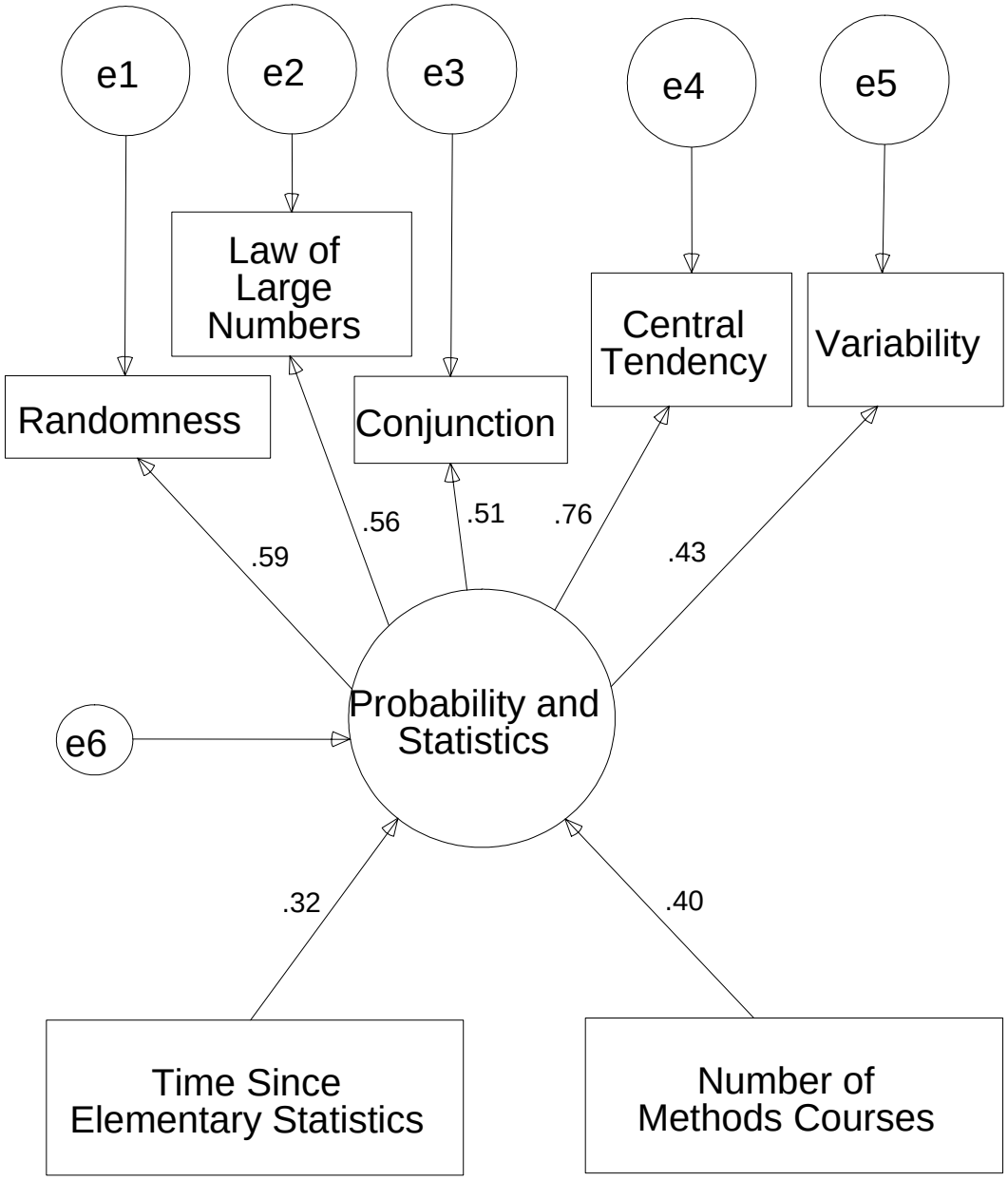


Figure 7. Single-Factor Time and Methods Model Standardized Regression Coefficients.

Table 16
Single-Factor Time and Methods Model Path Coefficients

Path	Unstandardized	Standard Error	Critical Ratio	P	Standardized
Probability and Statistics from Time Since Statistics	1.94	0.45	4.30	<0.001	0.32
Probability and Statistics from Number of Methods Courses	1.88	0.36	5.26	<0.001	0.40
Randomness from Probability and Statistics	0.33	0.05	6.91	<0.001	0.59
Conjunction from Probability and Statistics	0.54	0.09	6.15	<0.001	0.51
Central Tendency from Probability and Statistics	4.27	0.94	4.56	<0.001	0.76
Variability from Probability and Statistics	0.23	0.04	5.27	<0.001	0.43
Law of Large Numbers from Probability and Statistics	0.32	0.05	6.63	<0.001	0.56

Using a multivariate regression with the variables time since elementary statistics and number of methods courses regressed on the composite variables for likelihood and summary data, the overall model yielded $R^2 = 0.062$ for likelihood and $R^2 = 0.070$ for summary data (see Table 17). Time since elementary statistics did not contribute significantly to likelihood, but did have a statistically significant contribution on summary data. Number of methods courses made a statistically significant contribution to both likelihood and summary data. Although three of the four contributions were statistically significant, effect sizes were small.

Table 17
Multivariate Regression on Likelihood and Summary Data

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	P	Partial Eta Squared
Corrected Model	Likelihood	697[a]	2	348	6.80	.001	.062
	Summary	717[b]	2	359	7.83	.001	.070
Intercept	Likelihood	6906	1	6906	134.76	.000	.394
	Summary	6037	1	6037	131.80	.000	.389
Number of Methods Courses	Likelihood	664	1	664	12.96	.000	.059
	Summary	579	1	579	12.63	.000	.058
Time Since Elementary Statistics	Likelihood	161	1	161	3.14	.078	.015
	Summary	610	1	610	13.32	.000	.060
Error	Likelihood	10608	207	51			
	Summary	9481	207	46			
Total	Likelihood	234289	210				
	Summary	251736	210				
Corrected Total	Likelihood	11305	209				
	Summary	10198	209				

[a]R Squared = .062 (Adjusted R Squared = .053)
[b]R Squared = .070 (Adjusted R Squared = .061)

Because the extended-response items were phrased as simulated explanations to students, one would expect the methods courses to have a stronger impact on extended-response items than on multiple-choice items. In order to evaluate this assumption, a multivariate regression was used with the variables time since elementary statistics and number of methods courses regressed on the composite variables for non-extended-response and extended-response. The overall model yielded $R^2 = 0.045$ for non-extended-response and $R^2 = 0.101$ for extended-response (see Table 18). Time since elementary statistics did not contribute significantly to non-extended-response items, but did have a statistically significant contribution on extended-response. Number of methods courses made a statistically significant contribution to both non-extended and extended-response. Although three of the four contributions were statistically significant and effect sizes were small (Huck, 2004), both time since statistics and number of methods classes had a larger impact on extended-response than on non-extended-response. The largest effect was the methods courses on extended-response.

Table 18
Multivariate Regression on Non-Extended-Response and Extended-Response

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	P	Partial Eta Squared
Corrected Model	Non-Extended-Response	603[a]	2	302	4.93	.008	.045
	Extended-Response	704[b]	2	352	11.60	.000	.101
Intercept	Non-Extended-Response	14660	1	14660	239.45	.000	.536
	Extended-Response	1578	1	1578	52.04	.000	.201
Number of Methods Courses	Non-Extended-Response	600	1	600	9.80	.002	.045
	Extended-Response	642	1	642	21.18	.000	.093
Time Since Elementary Statistics	Non-Extended-Response	219	1	219	3.58	.060	.017
	Extended-Response	510	1	510	16.83	.000	.075
Error	Non-Extended-Response	12673	207	61			
	Extended-Response	6276	207	30			
Total	Non-Extended-Response	448757	210				
	Extended-Response	99254	210				
Corrected Total	Non-Extended-Response	13276	209				
	Extended-Response	6980	209				

[a]R Squared = .045 (Adjusted R Squared = .036)
[b]R Squared = .101 (Adjusted R Squared = .092)

CHAPTER V

CONCLUSIONS

Analysis of Assessment Responses

The first three research questions were based on the analysis of participants' responses to assessment items. Although the analyses were separated into multiple-choice and non-multiple-choice responses, both item types contribute to the overall analysis of preservice teachers' knowledge and understanding of probability and statistics, so they will be discussed jointly.

Although participants scored well on select Likelihood items, based on the descriptive statistics, participants generally answered Summary Data items (Central Tendency and Variability) correctly more often than Likelihood items (Randomness, Law of Large Numbers, and Conjunction). Considering that all of the participants had taken or were currently enrolled in a statistics class and did not have a probability class in their required degree plans, this result is not surprising. However, these participants will be expected to teach both probability and statistics concepts, so they should be thoroughly prepared to teach both areas when they graduate from their certification program.

Responses to Randomness

The INTASC Mathematics Subcommittee of the Council of Chief State School Officers was so concerned with the gambler's fallacy misconception among future teachers, that they included an example similar to item 12 in their standards document (1995). Therefore, it is heartening that very few participants (24 which is 11.4%) missed

this item. For the birth order item, 136 participants (64.7%) picked the correct answer choice and 80 participants (38.1%) provided either desired or acceptable explanations. This indicates the need to stress explanations rather than just short answers from preservice teachers.

Participants performed better on randomness than on the other Likelihood concepts, but it is important to note that the participants demonstrated an over-reliance on the belief that events are equally likely and the correct answer choice for both randomness items was that the events were equally likely. Therefore, it is unclear whether knowledge of randomness concepts or dependence on equally likely outcomes is the primary cause for this difference.

Responses to Law of Large Numbers

Only 39 participants (18.6%) provided an answer corresponding to the use of the law of large numbers on item 14, and even fewer (32, which is 15.2%) provided a desired or acceptable explanation. This shows that the students did not identify the need for the use of the law of large numbers. Responses to item 11 further illuminate the problem by directly addressing the change in sample size and asking the participants about the effects on the statistics representing the distribution. Participants realized that the mean was not affected by sample size, but the majority of the participants also expected the standard deviation to remain constant over various sample sizes. This error is consistent with their error on item 14 and indicates an underlying misconception in need of remediation (the effects of sample size). Additionally, 20 (9.5%) of the participants displayed the confounding effects that knowledge about ratios and

proportions had on this problem with their extended-response answers comparing the ratios of the choices rather than using the law of large numbers.

Responses to Conjunction

Of the conjunction items, participants performed the best (84, which is 40.0%, correct) on the computational item similar to those commonly found in an introductory probability and statistics class (item 16) even though it required more steps of logic than the other problems. Therefore, this gives cause to believe that classroom lessons can improve the preparedness of preservice teachers. However, the overall poor performance on conjunction items shows that this is a concept that needs more attention. Extended responses (item 17b) gave evidence that the improper invocation of the representativeness heuristic was partially to blame for preservice teachers' difficulty with conjunction items.

Responses to Central Tendency

Although constructing a data set to meet specific central tendency requirements necessitates more familiarity with the definitions than is required to compute the statistics from a given data set, results from item 7 indicate that the participants were generally successful at constructing a data set to meet specific requirements for mean, median, and mode. However, they were not as successful at deciphering which measure of typicality had been used to arrive at a given answer given a set of data in a bar graph (cf., results from item 10). Although 89.0% of the participants could identify the use of a measure of typicality in item 10, almost half of the participants (47.1%) were unable to interpret multiple views and could only identify the single measure of typicality. These

results indicate that these participants could benefit from viewing problems from multiple perspectives rather than a focus on problems that have a single correct solution. These multiple views will be essential when the participants are in the classroom attempting to create examples and pose questions that will help their students understand measures of central tendency. Use of the term *average*, even in a simulated teaching setting, rather than a more precise term such as mean, median, or mode by the participants is also an indication of their lack of understanding of the many measures of center.

Responses to Variability

Participants performed well on measures of variability, but it was surprising that more students could correctly answer a question about standard deviation (a high-school and collegiate topic) than about range (the first measure of variability encountered by students in junior high school. The item that inquired about range (9b) did require participants to glean information from a bar graph. Unfortunately, only 81.9% of the participants correctly answered item 9a that was designed to detect problems with graph reading ability. Students are expected to learn to read bar graphs around third grade and continue using that concept throughout their mathematical educations, but the percentage of students who correctly read the graph (item 9a), was similar to the percentage who correctly used a standard deviation to identify the data range in which one would expect 68% of the data to lie. The extended-response variability item on standard deviation was the extended-response item on which students had more rubric scores of three and above than for any other topic on the assessment (participants scored better on item 10a, but

that was offset by the other central tendency items). It is also worth noting that of the participants who correctly read the graph in 9a, only 69.1% correctly identified the range in 9b. It is likely that item 9b is another example of additional mathematical concepts (domain and range of a function when graphed in a coordinate plane) confounding previously learned ideas (a bar graph represents the frequencies of discrete data points). Evidence for confounding ideas is provided by the 25 participants (11.9%) who reported that the range of the data was nine indicating that they treated the bar graph as though it were displaying two-dimensional data (i.e., treating frequencies as though they represented the independent variable of a function).

Observations across Responses

Participants' responses were encouraging on selected items, but improvement is needed in most areas. It is encouraging that the items covered most intently in the elementary statistics classes are also the items on which participants scored better. This indicates that program alterations designed to combat the deficiencies evident from this assessment are likely to positively impact the preparation of PK-8 teachers.

Although it is generally necessary to know the correct answer before one can correctly explain that answer, it was disconcerting to see that the difference between extended- and non-extended-response scores was so enormous. All of these participants are preparing to be teachers. Therefore, the ability to explain concepts will be vital to their success and should be stressed in their academic preparation.

Comparison of Responses Based on Certification Level Pursued

It seems appropriate that those who are specifically preparing to teach mathematics and science would perform to a higher standard on a probability and statistics assessment due to the greater extent of their mathematics background and concentrated interest in mathematics and science. However, many primary schools in the state of Texas employ the generalist model, rather than the specialist model, of education. Therefore, this raises great concerns that the preservice teachers earning certification through the 4-8 Language Arts and Social Studies route will not be prepared to teach probability and statistics concepts if the need arises (i.e., in a self-contained classroom).

Effects of Cumulative Exposure

Although none of the effect sizes were large, statistically significant effects did exist in the expected direction. The most notable effect was that students who had taken more methods classes provided better explanations even though the methods classes did not specifically cover much of the content tested. However, most of these students had taken no more than one methods class. A longitudinal study would be helpful to explore this effect further.

Issues for Further Investigation

Based on the following response to item 13b, an item should be added to determine if students interpreted item 13b as three of each gender vs. five of one gender and one of another rather than particular birth orders.

> It is equally likely that the mother will have any one of these combinations of children because the specific combinations given can only happen once. IF you were asking for the probability of getting just three boys and three girls in any order versus getting five girls and one boy in any order then the probability would be different.

One participant indicated that the word “carrying” in item 17 could have multiple interpretations.

> When you hear that a woman is carrying a baby you normally think of a woman being pregnant, but when you see the choices that she is a doctor she could have delivered a baby and carrying it or she could also be a mother carrying her own child.

Although the conjunction fallacy would be tested and same logic would be used to answer the question for both interpretations, the word “carrying” should be changed to “holding” to increase the likelihood of a more consistent interpretation of the situation.

The relationship between number of methods classes and extended responses surfaced in this study, but the majority of the students had taken no more than one methods course. Therefore, a longitudinal study would be helpful to explore the relationship between methods classes and quality of extended-response answers.

This study evaluated the preparedness of preservice teachers by their answers to written questions. However, this is only a proxy measure to discern what they are capable of teaching. Therefore, a study examining actual lessons taught by preservice teachers would be an appropriate extension to this study.

This study identified some of the strengths and weaknesses of the students emerging from this program. The next step in ensuring that the preservice teachers are prepared to teach probability and statistics is to conduct an assessment of the existing program to determine what changes can be made that would retain the positive benefits of the current program while strengthening students' preparation in the areas identified as weaknesses.

Concluding Remarks

The results of this survey indicate that improvements are necessary to ensure that PK-8 preservice teaches are prepared to teach probability and statistics. The NCTM states that students as young as third grade should start working with bar graphs, and students as young as sixth should be able to describe data sets using statistics such as median and range, yet almost a fifth of the preservice teachers in this study missed the items pertaining to these topics. Additionally, conceptual errors in statistical and probabilistic reasoning were common. The extended responses on this assessment indicate a gap between understanding and explaining a concept. Results of this assessment indicate that many preservice teachers are not yet prepared to teach probability and statistics with the deep understanding necessary for conceptual learning, so program changes are necessary if we want our preservice teachers to have a thorough understanding of probability and statistics concepts and to be able to explain those ideas to their students.

REFERENCES

The American heritage dictionary of the English language (4th ed.). (2000). Boston: Houghton Mifflin Company. Retrieved January 11, 2005, from www.dictionary.com

Beser, S. (2004). *Educational survey tool (Beta version)*. [Computer software]. Ankara, Turkey: Middle East Technical University.

Byrnes, J. P. (2001). *Cognitive development and learning in instructional contexts* (2nd ed.). Boston: Allyn and Bacon.

Carpenter, T. P., & Hiebert, J. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.). *Handbook of research on mathematics teaching and learning* (pp. 65-97). Reston, VA: National Council of Teachers of Mathematics.

Conference Board of the Mathematical Sciences. (2001). *The mathematical education of teachers*. Providence, RI: American Mathematical Society and Mathematical Association of America. Retrieved May 12, 2005 from http://www.cbmsweb.org/MET_Document/index.htm

Davidson, D. (1995). The representativeness heuristic and the conjunction fallacy effect in children's decision making. *Merrill-Palmer Quarterly, 41*, 328-346.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2000). *Handbook of qualitative research* (2nd ed). Thousand Oaks, CA: Sage.

Fast, G. R. (1997). Using analogies to overcome student teachers' probability misconceptions. *Journal of Mathematical Behavior, 16*, 325-344.

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

Gavanski, I., & Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology, 61,* 181-194.

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

Hopkins, K. D., & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement, 50,* 717-729.

Huck, S. W. (2004). *Reading statistics and research* (4th ed.). Boston: Pearson.

Interstate New Teacher Assessment and Support Consortium. (1995). *Model standards for beginning teacher licensing and development: A resource for state dialogue*. Washington, DC: Author.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3*(1). Retrieved June 5, 2004, from http://www.amstat.org/publications/jse/v3n1/konold.html

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

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Mertens, D. M. (2005). *Research and evaluation in education and psychology* (2nd ed.). Thousand Oaks, CA: Sage.

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

Mokros, J., & Russell, S. J. (1995). Children's concept of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20-39.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants* (pp. 95-137). Washington, D.C.: National Academy Press.

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

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

O'Connell, A. A. (1999). Understanding the nature of errors in probability problem-solving. *Educational Research and Evaluations, 5*, 1-21.

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept of computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*, 191-204.

Rubin, A., & Rosebery, A. (1988). Teachers' misunderstandings in statistical reasoning; Evidence from a field test of innovative materials. In A. Hawkins (Ed.), *Training teachers to teach statistics: Proceedings of the International Statistics Institute Roundtable Conference* (pp. 72-89). Voorburg, The Netherlands: International Statistical Institute.

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

Shaughnessy, J. M. (1981). Misconceptions of probability: From systematic errors to systematic experiments and decisions. In A. P. Schulte & J. R. Smart (Eds.), *NCTM 1981 yearbook* (pp. 90-100). Reston, VA: NCTM.

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

Sheskin, D. J. (2004). *Handbook of parametric and non-parametric statistical procedures* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, D.C.: American Psychological Association.

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

Truran, J. (2001). Postscript: Researching stochastic understanding-The place of a developing research field in PME. *Educational Studies in Mathematics, 45,* 9-13.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 4*, 293-315.

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

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

Watson, J. M., & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning, 2*(1 & 2), 11-50.

APPENDIX A

PROBABILITY AND STATISTICS SURVEY 2004

If you do NOT wish to participate, quit this survey now and do not submit it. No information will be recorded. No adverse actions will be taken against you or your grades if you choose this option. You will still participate in all the same tests, assignments, and other classroom activities as the rest of the class.

By submitting this survey, you are agreeing to the consent form from the previous page. (If you would like to read it again, please quit this survey and start again. There is a link to it on the first screen. You can also print the consent form from that screen).

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study. I have been given a copy of this consent form (you can print the consent form from the link on the previous page).

Any questions can be directed to Tamara Carter.

A calculator is not necessary, but it might be helpful. You are welcome to use a calculator on this survey, but please do not use any other resources (including other people) - the results of this survey will NOT average into your class grade. We just want to know what you remember off the top of your head.
Your answers will not be recorded until you submit the survey.

Thank you for your time and effort! We really appreciate your help!

Item 1: What is your gender?

- Male
- Female

Item 2: What is your current class level?

- Freshman
- Sophomore
- First Semester Junior
- Second Semester Junior
- First Semester Senior
- Second (or more) Semester Senior
- Graduate Student
- Other

Item 3: Which of the following best describes your major?

- Education - Early Childhood
- Education - Gr. 4-8 Language Arts / Social Studies
- Education - Gr. 4-8 Math / Science
- Agriculture / Architecture / Science WITH teaching certificate

- Agriculture / Architecture / Science WITHOUT teaching certificate
- Business WITH teaching certificate
- Business WITHOUT teaching certificate
- Engineering WITH teaching certificate
- Engineering WITHOUT teaching certificate
- Liberal Arts WITH teaching certificate
- Liberal Arts WITHOUT teaching certificate
- Other

Item 4: What is your ethnic background?

- American Indian or Alaskan Native
- Asian or Pacific Islander
- African American
- Hispanic / Latino
- White (non-Hispanic)
- Other or prefer not to answer

Item 5: For each of the following courses, indicate which semester you took the course (if at all).

	Did not take this course	Prior to Fall 00	Fall 2000	Spr 01	Sum 01	Fall 2001	Spr 02	Sum 02	Fall 2002	Spr 03	Sum 03	Fall 2003	Spr 04	Sum 04	Fall 2004 (this semester)
ECFB 440															
EPSY 435															
MASC 341															
MASC 450															
MEFB 450															
MEFB 460															
STAT 303															

Item 6: For the course(s) listed below that you are taking **this semester**, please indicate your instructor(s).

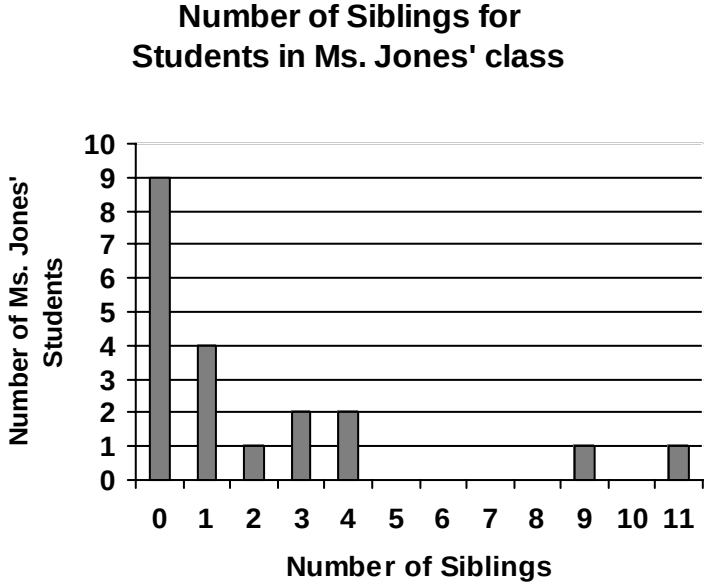
- ECFB 440 with Instructor A
- ECFB 440 with Instructor B
- ECFB 440 with Instructor C
- EPSY 435 with Instructor D
- EPSY 435 with Instructor E
- EPSY 435 with Instructor F
- MASC 351 with Instructor G
- MASC 351 with Instructor H
- MASC 450 with Instructor J
- MEFB 450 with Instructor T
- MEFB 450 with Instructor K
- MEFB 460 with Instructor U
- STAT 303 with Instructor L
- STAT 303 with Instructor M
- STAT 303 with Instructor N
- STAT 303 with Instructor P
- STAT 303 with Instructor Q
- STAT 303 with Instructor R
- STAT 303 with Instructor S

Item 7: Find five real numbers so that the mean is 85, the median is 81, and the mode is 78. Put one number in each blank.
First Number:
Second Number:
Third Number:
Fourth Number:
Fifth Number:

Item 8: On a national exam, the scores were normally distributed with a mean of 60 and standard deviation of 12. You would expect approximately 68% of the scores to fall between which two numbers?

- 54 and 66
- 48 and 72
- 36 and 84
- 24 and 96

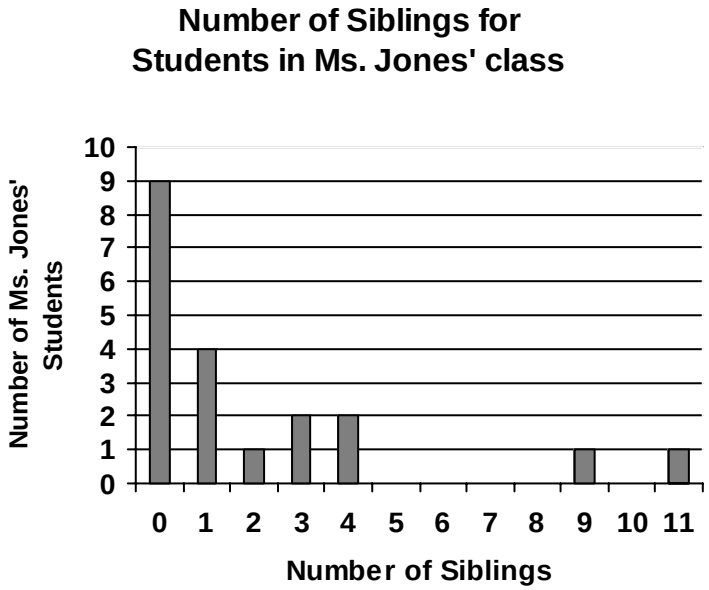
Item 9: Please answer the questions below using this graph.



9a: How many of Ms. Jones' students have three or more siblings?

9b: What is the range of the data?

Item 10: You gave this graph (depicting the number of siblings for each of Ms. Jones' 20 students) to a 7th grade class and asked, "How many siblings does the typical student in Ms. Jones' class have?" Shelly, Jennifer, and Henry all had different answers, but they all used correct computational methods to give you an exact answer. **Please explain how each student interpreted the word "typical" and explain how he/she arrived at his/her answer.**



10a: Shelly said 0 siblings.

10b: Jennifer said 1 sibling.

10c: Henry said 2 siblings.

10d: For this data set, which meaning of the word typical would you use? Please explain your reasons for this choice.

Item 11: Students at Typical Junior High were randomly placed into the twelve seventh-grade math classes. There are 30 students in each class. Using all 360 scores, calculations revealed a mean score of 75 and a standard deviation of 11.

11a: Would you expect the mean of an individual class to be (less than, more than, or approximately equal to) 75?

11b: Would you expect the standard deviation of an individual class to be (less than, more than, or approximately equal to) 11?

11c: One of the students in this class said she had never heard of a standard deviation. How would you explain standard deviation to this seventh-grade student without using the formula?

11d: How can the standard deviation be used to describe the distribution of the 360 scores in relation to the mean?

Item 12: You have flipped a **FAIR** coin 10 times and it landed on heads all 10 times. Which of the following is more likely to occur on the next flip?

- Coin lands on heads
- Coin lands on tails
- Both choices (landing on heads and landing on tails) are equally likely

Item 13a: A family has six children. Which of the following choices best describes the order in which the children are most likely to be born?

- GGGBBB
- BGGBBG
- GBGBGB
- GGGGGB
- The first three choices given are equally likely and more likely than the fourth choice.
- The first four choices given are all equally likely.

13b: Explain your reasoning in detail as if you were explaining to a 7th grade student.

Item 14a: The principal in a junior high school believes in bringing math to life, so she offered the following incentive to each 7th grade class.:
"I will flip a fair coin the number of times that you pick - no matter how long it takes. If the coin lands on tails at least 60% of the time, you can have a pizza party on Friday. How many times do you want me to flip the coin?"
Casey wants lots of flips for the best chance at a party and picks 5000 flips.
Kelly says only flipping 5 times would give a better chance.
Shannon said that it doesn't matter how many times the coin is flipped. You have the same chance of a party with 5 flips as you do with 5000 flips.
Which student is correct?

- Casey
- Kelly
- Shannon

14b: Explain to these students in detail why this student is correct.

Item 15: A seventh-grade teacher in a school introduced his probability unit by giving his students a game involving one red die and one blue die. In the game, he would roll both dice once. The task for the students was to decide which of the following options would give them the best chance of winning this game. Which would you pick?

- You roll a 5 on the red die.
- You roll a 5 on the red die, and a 6 on the blue die.
- You roll a 5 on the red die, and a number other than 6 on the blue die.
- It doesn't matter. All of the previous three choices give the same chance of winning.

Item 16: You discovered a specially weighted coin that has a probability of 0.6 of landing on **TAILS** for any particular flip. If you flip this coin twice, what is the probability that it will land on **HEADS** both times?

- 0
- 0.16
- 0.25
- 0.36
- 0.4
- 0.5
- 0.6
- 0.8
- 1
- 1.2

Item 17a: You see a woman carrying a baby. Which of the following is more likely?

- The woman is a doctor.
- The woman is a doctor and a mother.
- Both of these choices are equally likely.

17b: Explain your reasoning in detail as if you were explaining to a 7th grade student.

APPENDIX B

INSTRUCTOR SURVEY

Item 1
Thank you VERY much for helping me with this survey!!!!!
I would appreciate it if you would take the time to answer a few questions yourself.
During the time this survey was administered, **how many** students were enrolled in your class(es)? (I am asking about the class(es) that we asked to participate in this survey).

Item 2
Please describe your communications with your students about this survey.
Did you post it on a website?
. . . tell them in class (If so, how often did you mention it)?
. . . email reminders (If so, the ones I sent, or others)?

Item 3
Did your students have any incentive (other than helping me) to complete the survey? If so, what?

Items 4-13
The questions that follow are the questions that were on the survey that your students took. The bulleted items are the choices for multiple-choice questions.
I am **NOT** asking you to answer the questions. I would like to know your comments on the questions.
For each question . . .
Was the item covered in your class prior to the survey? (I realize that there are many questions that are not supposed to be covered in your class.)
Is there a particular reason (wording or something) that the question might cause students a problem?
Did students make any interesting comments about the question?
Do you have any other thoughts about the question?

APPENDIX C

PARTICIPANT CONSENT FORM FOR OBSERVATIONS, INFORMATION RELEASE, AND INTERVIEWS

Statistics Education of Future K-8 Texas Teachers

The purpose of the study:

I understand that the purpose of this study is to understand more about what future K-8 teachers do and do not know about probability and statistics. Since many students who take this class are preparing to teach, I have been asked to participate regardless of my major. This is not an experiment. The researcher will not attempt to change the manner in which this class is taught.

I agree to the following during Fall 2004.

1. My instructor may provide information to the researcher including my grades from this class, samples of my work from this class, my age, gender, major, and classification (Freshman, Sophomore, Junior or Senior).
2. The researcher may request to speak with me about my understanding of specific statistical concepts and my attitudes toward teaching and statistics. I can accept or decline this invitation without repercussions and still participate in other parts of the study.

****

I understand that:

1. Participation is strictly voluntary. I can refuse to answer any questions that I do not wish to answer.
2. The information gathered will not affect grades or any other evaluations made by the teacher of this course.
3. The information gathered will be confidential. Student and teacher names or any other identifying factors will be removed from any report or publication of the data or results.
4. I may opt out of the project at any time and for any reason I deem necessary with no repercussions if I give written notice to the researcher.
5. Approximately 700 students per semester in certain sections of STAT 303, EPSY 435, MASC 351, MASC 450, ECFB 440, MEFB 450 and MEFB 460 have been asked to participate.
6. Participation in this study will not directly provide any benefits to me. Declining participation in this study will not cause adverse actions to be taken against me or my grades.
7. The researcher will observe some class sessions during the semester but will not audio or video tape the classes.

I understand that this research study has been reviewed and approved by the Institutional Review Board –Human Subjects in Research, Texas A&M University. For research-related problems or questions regarding subjects' rights, I can contact the Institutional Review Board through Dr. Michael W. Buckley, Director of Research Compliance, Office of Vice President for Research at (979) 458-4067 (mwbuckley@tamu.edu).

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study. I have been given a copy of this consent form.

Student's name PRINTED _____

Student's Signature _____ Date_____

Researcher's Signature _____ Date _____

If I do NOT wish to participate I will not return this form. No adverse actions will be taken against me or my grades if I choose this option. I will still participate in all the same tests, assignments, and other classroom activities as the rest of the class.

If you have any questions or concerns, please contact:

Researcher: Tamara Carter
TLAC Ph.D. student, Texas A&M University, MS 4232, College Station, TX 77843-4232, (979) 458-3888.

Student of: Dr. Gerald Kulm, Curtis D. Robert Professor
TLAC Dept., Texas A & M University, MS 4232, College Station, TX 77843-4232, (979) 862-4407.

APPENDIX D

INSTRUCTOR CONSENT FORM FOR OBSERVATIONS, INFORMATION RELEASE, AND INTERVIEWS

Statistics Education of Future K-8 Texas Teachers

The purpose of the study:

I understand that the purpose of this study is to understand more about what future K-8 teachers do and do not know about probability and statistics. Since many students who take this class are preparing to teach, all of my students will be asked to participate regardless of their major. This is not an experiment.

I agree to the following during Fall 2004.

1. The researcher will not attempt to change the manner in which this class is taught.
2. I will provide information to the researcher about the students participating in this study including age, gender, major, classification (Freshman, Sophomore, Junior or Senior), grades from this class, and samples of work from the students (homework, tests, etc.) from this class.
3. I will provide the researcher with access to my syllabus for this class and any other material that is available to the students in this class.
4. The researcher may request to speak with me about specific statistical concepts, my attitudes toward teaching and statistics, and my perceptions of this particular class. I can accept or decline this invitation without repercussions and still participate in other parts of the study.
5. The researcher may request to speak with some of my students about their understanding of specific statistical concepts and their attitudes toward teaching and statistics. They can accept or decline this invitation without repercussions and still participate in other parts of the study.
6. The researcher may ask me to give a statistics assessment to my class that will be similar to a quiz that might be given in a concepts oriented statistics class. I can accept or decline this invitation without repercussions and still participate in other parts of the study.

****

I understand that:

1. Participation is strictly voluntary. I can refuse to answer any questions that I do not wish to answer.
2. I will not coerce my students to participate or not to participate in this study.
3. The information gathered will not affect my students' grades or any other evaluations made by the teacher of this course.
4. The information gathered will not affect my professional evaluations in any manner.

5. The information gathered will be confidential. Student and teacher names or any other identifying factors will be removed from any report or publication of the data or results.
6. I may opt out of the project at any time and for any reason I deem necessary with no repercussions if I give written notice to the researcher.
7. Approximately 700 students per semester in certain sections of STAT 303, EPSY 435, MASC 351, MASC 450, ECFB 440, MEFB 450 and MEFB 460 have been asked to participate.
8. Participation in this study will not directly provide any benefits to me. Declining participation in this study will not cause adverse actions to be taken against me.
9. The researcher will observe some class sessions during the semester but will not audio or video tape the classes.

I understand that this research study has been reviewed and approved by the Institutional Review Board –Human Subjects in Research, Texas A&M University. For research-related problems or questions regarding subjects' rights, I can contact the Institutional Review Board through Dr. Michael W. Buckley, Director of Research Compliance, Office of Vice President for Research at (979) 458-4067 (mwbuckley@tamu.edu).

I have read and understand the explanation provided to me. I have had all my questions answered to my satisfaction, and I voluntarily agree to participate in this study. I have been given a copy of this consent form.

Instructor's name PRINTED _____

Instructor's Signature _____Date_____

Researcher's Signature _____Date _____

I have read and understand the explanation provided to me, but I do NOT wish to participate. By printing my name in the space below, I am indicating that I do not wish to participate so that the researcher will not attempt to contact me again about this study. NO adverse actions will be taken against me for choosing this option.

If you have any questions or concerns, please contact:
Researcher: Tamara Carter
TLAC Ph.D. student, Texas A&M University, MS 4232, College Station, TX 77843-4232, (979) 458-3888.

Student of: Dr. Gerald Kulm, Curtis D. Robert Professor
TLAC Dept., Texas A & M University, MS 4232, College Station, TX 77843-4232, (979) 862-4407.

APPENDIX E

COURSE DESCRIPTIONS

ECFB 440. Mathematics Methods in Early Childhood Education. (2-6). Credit 3.
Analyzes contemporary curricula; implementation of methods relevant for active, authentic learning and age appropriate teaching of mathematics to young learners; considers state and national standards related to teaching and learning mathematics. Prerequisites: ECHE 332 and 342; admission to teacher education; senior classification; Corequisites: ECFB 400 and 420; RDNG 440.

EPSY 435. Educational Statistics. (3-0). Credit 3.
Statistical concepts and techniques and their application in behavioral sciences. Prerequisite: Junior or senior classification.

MASC 351. Problem Solving in Mathematics. (3-0). Credit 3.
Problem solving strategies in math and science; evaluate conjectures and arguments; writing and collaborating on problem solutions; posing problems and conjectures; constructing knowledge from data; developing relationships from empirical evidence; connecting mathematics concepts; readings, discussions, and analyses will model and illustrate mathematics problems solving and proofs. Prerequisite: 9 hours of 300-level mathematics courses; admission to teacher education; junior classification.

MASC 450. Integrated Mathematics. (3-0). Credit 3.
Integration and connections among topics and ideas in mathematics and other disciplines; connections between algebra and geometry and statistics and probability; focus for integration with authentic problems requiring various branches of mathematics. Prerequisites: MASC 351; admission to teacher education; junior classification.

MEFB 450. Social Studies Methods in the Middle Grades. (2-6). Credit 3.
Trends and issues related to middle grades curriculum development and instruction in social studies and humanities; integration of content, planning, teaching-learning experiences; evaluation of teaching and learning in social studies. Prerequisites: MEFB 352; MIDG 352; admission to teacher education; senior classification; Corequisites: MEFB 480 and 490; RDNG 470 and 490.

MEFB 460 Math Methods in Middle Grades. (2-6). Credit 3.
Examines theories, provides practice in teaching methods essential to successful mathematics learning; focuses on content and criteria central to teaching mathematics for understanding, skill development, and problem solving; readings, discussions, analyses; modeling and practicing mathematics teaching and learning. Prerequisites: MEFB 352; MIDG 352; admission to teacher education; senior classification; Corequisites: MEFB 470, 480, 490; MASC 450.

STAT 303 Statistical Methods. (3-0). Credit 3.
Intended for undergraduate students in the social sciences. Introduction to concepts of random sampling and statistical inference, estimation and testing hypotheses of means and variances, analysis of variance, regression analysis, chi-square tests. Credit will not be allowed for more than one of STAT 301, 302 or 303. Prerequisite: MATH 141 or 166 or equivalent.

APPENDIX F

SURVEY PARTICIPATION

Course	Instructor	Incentives	Number of students enrolled in Class(es)[a]	Number who took assessment[a]	Percentage of class participating	Number in Analysis[a]
ECFB 440	A	Time in class to complete assessment	22	21	95%	19
ECFB 440	B	None	26 & 26	32	61%	30
ECFB 440	C	None	56	10	18%	9
EPSY 435	D	Half a point on semester grade	50	39	78%	23
EPSY 435	E	None	50	9	18%	5
EPSY 435	F	Two thirds of a point on semester grade	76	60	79%	47
MASC 351	G	One point on semester grade	30	19	63%	14
MASC 351	H	None	27	1	4%	1
MASC 450	J	1.27 points on semester grade	33	20	61%	20
MEFB 450	K	Replace one Chapter Reflection	40	27	68%	25
STAT 303	L	Extra Optional Quiz grade of 100. Quiz average is 5% of final average.	47	43	91%	6
STAT 303	M	None	50, 50, & 49	15	10%	9
STAT 303	N	Extra Optional Quiz grade of 100. Quiz average is 5% of final	47	28	57%	4
STAT 303	P	None	45 & 45	3	3%	1
STAT 303	Q	None	39	7	18%	1
STAT 303	R	None	49	4	8%	2
STAT 303	S	None	50	30	60%	9

[a]Some students were enrolled in more than one of these classes during FA04, so they would be counted twice in this chart.

APPENDIX G

RUBRIC FOR DISSERTATION DATA

When coding, if answer contains parts of multiple codes, use the highest one that fits appropriately. For the extended response questions, it is particularly tempting to judge answers by the answers to multiple-choice questions. However, extended response answers should be judged independently of the other work unless the answer references previous work.

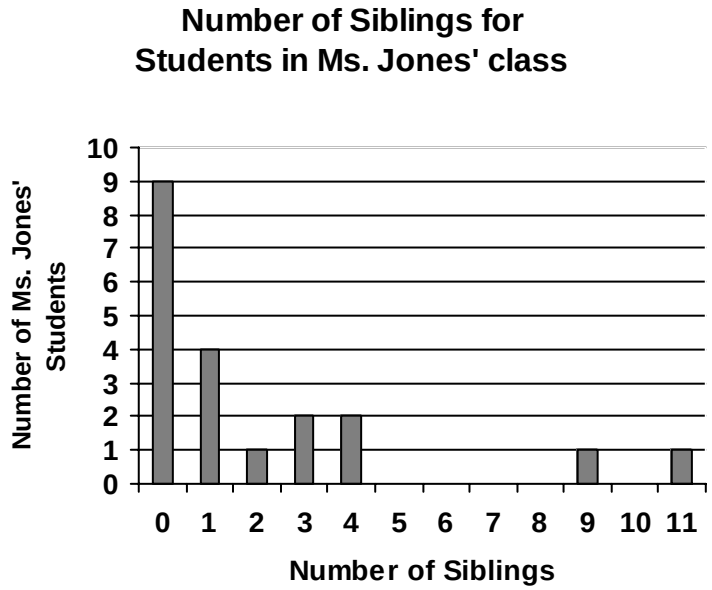
Item 7 (short-answer): Find five real numbers so that the mean is 85, the median is 81, and the mode is 78. Put one number in each blank.
First Number:
Second Number:
Third Number:
Fourth Number:
Fifth Number:

- 7a Mean
  - 5 = Mean of the 5 numbers is 85 (i.e., the sum is 425),
  - 1 = Mean is not 85
- 7b Median
  - 5 = Median is 81,
  - 3 = 81 was the middle number listed but is not middle when arranged smallest to largest,
  - 1 = Median is Not 81 and 81 was not middle number listed
- 7c Mode
  - 5 = 78 is mode with 2,
  - 4 = 78 is one of 2 modes,
  - 3 = 78 is one of 5 modes (all 5 numbers different),
  - 2 = 78 is not the mode but 78 is used,
  - 1 = 78 is not used

Item 8 (multiple-choice): On a national exam, the scores were normally distributed with a mean of 60 and standard deviation of 12. You would expect approximately 68% of the scores to fall between which two numbers?

- 5 = 48 and 72,
- 3 = 54 and 66,
- 2 = 36 and 84,
- 1 = 24 and 96

Item 9: Please answer the questions below using this graph.



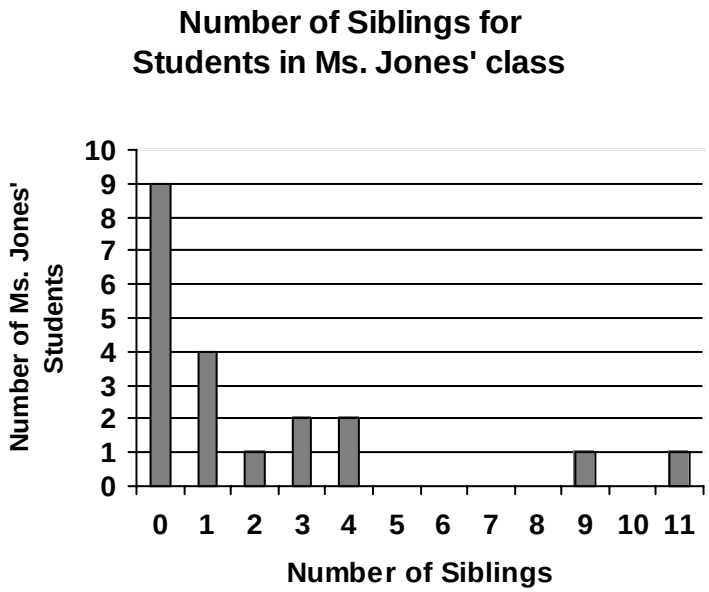
9a (short-answer): How many of Ms. Jones' students have three or more siblings?

- 5 = Six
- 3 = Four
- 1 = other answers
- 0 = No answer

9b (short-answer): What is the range of the data?

- 5 = for 11, “0 to 11”, or the discrete range, "0,1,2,3,4,9,11"
- 3 = for 9 , "0 to 9", or the discrete range "0,1,2,4,9"
- 1 = other answers
- 0 = No answer

Item 10: You gave this graph (depicting the number of siblings for each of Ms. Jones' 20 students) to a 7th grade class and asked, "How many siblings does the typical student in Ms. Jones' class have?" Shelly, Jennifer, and Henry all had different answers, but they all used correct computational methods to give you an exact answer. **Please explain how each student interpreted the word "typical" and explain how he/she arrived at his/her answer.**



10a (extended response): Shelly said 0 siblings.

- 5 = Desired Answer: Used the word “mode” AND explained the meaning of mode.
- 4 = Acceptable Answer: Used the word “mode” OR explained the meaning of mode.
- 3 = Incomplete Answer: Vague explanation or referred only to the appearance of the graph without an interpretation of its meaning OR explained how Shelly could have misread the graph to achieve the answer of 0.
- 2 = Incorrect Answer: Indicated mean or median, read graph incorrectly, or was incorrect in another manner.
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as “don’t know” rather than attempting an answer to the question

10b: (extended response) Jennifer said 1 sibling.

- 5 = Desired Answer: Used the word “median” AND explained the meaning of median
- 4 = Acceptable Answer: Used the word “median” OR explained the meaning of median OR computed the mean after removing the outliers.
- 3 = Incomplete Answer: Vague explanation OR an error in an explanation of median OR explained how Jennifer could have misread the graph to achieve the answer of 1.
- 2 = Incorrect Answer: Indicated mean or mode, read graph incorrectly, or was incorrect in another manner.

- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as “don’t know” rather than attempting an answer to the question

10c: (extended response) Henry said 2 siblings.

- 5 = Desired Answer: Used the word “mean” AND explained the meaning of mean
- 4 = Acceptable Answer: Used the word “mean” OR explained the meaning of mean OR used the word “average” without mentioning which definition of average was being used OR indicated that the mean must be rounded to yield an answer of 2.
- 3 = Incomplete Answer: Vague explanation OR explained how Henry could have misread graph to achieve this answer.
- 2 = Incorrect Answer: Indicated mode or median, read graph incorrectly, or was incorrect in another manner.
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as “don’t know” rather than attempting an answer to the question

10d: (extended response) For this data set, which meaning of the word typical would you use? Please explain your reasons for this choice.
NOTE: The choice of typicality is coded in 10e and has no correct answer. The explanation of the choice ONLY is coded in 10d.

- 5 = Desired Answer: Provided an accurate, well-stated, statistically-based reason (other than the definition of the choice) that matched the choice. Possible reasons could mention outliers, normality, skewness, probability of correctness, or other factors.
- 4 = Acceptable Answer: Provided a reason that matched the choice. Possible reasons could mention outliers, normality, skewness, probability of correctness, or other factors.
- 3 = Incomplete Answer: The reason given did not match the choice of typicality or the reason given was that the chosen measure was the most accurate or effective choice (without stating other reasons).
- 2 = Incorrect Answer: Only described the choice of typicality without providing a reason for the choice, restated the choice as the explanation (e.g., I would pick 1 because it is the median), or did not explain why the choice was a good measure of typicality.
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as “don’t know” rather than attempting an answer to the question

10e (extended response) Typical Type

NOTE: there is NO CORRECT ANSWER TO THIS QUESTION, so the numbers attached to these codes are arbitrary, and this problem is not used in the composite score for the instrument.

- 5 = Mode: Used the word "mode", the concept of mode, or agreed with 0, Shelly, or first choice.
- 4 = Median: Used the word "median", the concept of median, or agreed with 1, Jennifer, or the second choice.
- 3 = Mean: Used the word "mean", the concept of mean, or agreed with 2, Henry, or the third choice.
- 2 = Average (undefined): Use the word average without indication of which meaning of average, or described multiple measures of center without picking one or used words and descriptions that did not reflect the same measure of center.
- 1 = Provided an answer that is not a measure of center or is otherwise incorrect
- 0 = Participant did not answer the question or made a comment such as "don't know" rather than attempting an answer to the question

Item 11: Students at Typical Junior High were randomly placed into the twelve seventh-grade math classes. There are 30 students in each class. Using all 360 scores, calculations revealed a mean score of 75 and a standard deviation of 11.

11a: (multiple-choice) Would you expect the mean of an individual class to be (less than, more than, or approximately equal to) 75?

- 5 = Approximately Equal
- 2 = Less than
- 1 = More than
- 0 = Other answer

11b: (multiple-choice) Would you expect the standard deviation of an individual class to be (less than, more than, or approximately equal to) 11?

- 5 = More Than
- 3 = Less Than
- 1 = Approximately Equal
- 0 = Other answer

11c: (extended response) One of the students in this class said she had never heard of a standard deviation. How would you explain standard deviation to this seventh-grade student without using the formula?

- 5 = Desired Answer: Standard deviation is the amount that the data vary around the mean **on average** or another appropriate definition
- 4 = Acceptable Answer: Mentions that the variability is **around the mean** but does not explain further or had other errors in the explanation (in this instance, accept the word "average" in place of "mean")

- 3 = Incomplete Answer: Explains “deviation” rather than “standard deviation” or uses ideas related to the percentage of the data bounded by standard deviations under a normal curve (68%, 95%, and 99%).
- 2 = Incorrect Answer: Used the word “deviate” as an important part of the explanation, gives an that does not help explain standard deviation beyond the formula, or is incorrect in another manner.
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as “don’t know” rather than attempting an answer to the question

11d: (extended response) How can the standard deviation be used to describe the distribution of the 360 scores in relation to the mean?

- 5 = Desired Answer: If data are distributed **normally**, approximately **68**% of the data falls within 1 STD of the mean, and approximately **95**% falls within 2 STD. (Figures that are close to 68% and 95% with or without the word *approximately* are acceptable. This category can still be used even if normality is not mentioned.)
- 4 = Acceptable Answer: Approximately 68% of the scores are between 64 and 86 (or one STD from the mean) OR 95% within 2 STD OR 99% within 3 STD (or similar numbers) OR an answer that mentions more than one of these options but contains an error.
- 3 = Incomplete Answer: Most of the data (does not give percentage or percentage is incorrect) are between 64 and 86 (or one STD from the mean) OR mentions 64 and 86 (or add and subtract STD from mean) but not say why those numbers are important OR gives the percentages 68%, 95%, and 99% without explaining how they are related to standard deviation OR references z-scores or the empirical rule without explaining their connection to standard deviation OR defines standard deviation OR makes the connection between the magnitude of the standard deviation and the spread of the scores.
- 2 = Incorrect Answer: Used the word “deviate” as an important part of the explanation, gives an that does not help explain the distribution of data, expects all the data to fall within the first standard deviation, was significantly off on the percentages, or was incorrect in another manner.
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as “don’t know” rather than attempting an answer to the question

11e: (combination of 11c and 11d for cumulative exam score)

- 5 = Desired Answer: Earned a 4 or 5 on BOTH 11c and 11d.
- 4 = Acceptable Answer: Earned a 4 or 5 on 11c OR 11d.
- 3 = Incomplete Answer: Max(11c and 11d) = 3
- 2 = Incorrect Answer: Max(11c and 11d) = 2

- 1 = Max(11c and 11d) = 1
- 0 = Earned a 0 on both 11c and 11d.

Item 12: (multiple-choice) You have flipped a **FAIR** coin 10 times and it landed on heads all 10 times.
Which of the following is more likely to occur on the next flip?

- 5 = Both choices (landing on heads and landing on tails) are equally likely
- 2 = Coin lands on tails
- 1 = Coin lands on heads

Item 13a: (multiple-choice) A family has six children. Which of the following choices best describes the order in which the children are most likely to be born?

- 5 = The first four choices given are all equally likely
- 4 = The first three choices given are equally likely and more likely than the fourth choice
- 3 = BGBGBG
- 2 = "GGGBBB" or "BGGBBG"
- 1 = GGGGGB

13b: (extended response) Explain your reasoning in detail as if you were explaining to a 7th grade student.

- 5 = Desired Answer: Used the word INDEPENDENT to describe individual births or said that previous births do not affect the next birth.
- 4 = Acceptable Answer: 50/50 chance EACH birth (without adding extra mathematically incorrect information)
- 3 = Incomplete Answer: 50/50 chance (not mention that chance is same for EACH birth) OR an answer that would have earned a 5 or 4 but had extra mathematically incorrect information OR uses extra information (such as saying ALL birth orders are equally likely) or another method of explanation to demonstrate that the answer was more than a guess.
- 2 = Incorrect Answer: Context driven answer (representativeness such as "based on the families I have seen . . ." or "genetics determines . . .") OR expect an equal number of each gender OR say it is all based on chance OR incorrect in another manner.
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as "don't know" rather than attempting an answer to the question

Item 14a: (multiple-choice) The principal in a junior high school believes in bringing math to life, so she offered the following incentive to each 7th grade class.:

"I will flip a fair coin the number of times that you pick - no matter how long it takes. If the coin lands on tails at least 60% of the time, you can have a pizza party on Friday. How many times do you want me to flip the coin?"
Casey wants lots of flips for the best chance at a party and picks 5000 flips.
Kelly says only flipping 5 times would give a better chance.
Shannon said that it doesn't matter how many times the coin is flipped. You have the same chance of a party with 5 flips as you do with 5000 flips.
Which student is correct?

- 5 = Kelly (5 flips)
- 3 = Shannon (Doesn't matter)
- 1 = Casey (5000 flips)

Item 14b: (extended-response) Explain to these students in detail why this student is correct.

- 5 = Desired Answer: Explains the idea of the law of large numbers (more flips are more likely to approximate the theoretical probabilities of 50/50, so fewer flips gives a better chance for the party) and related it to the specific numbers or concepts in the problem
- 4 = Acceptable Answer: Law of large numbers words or idea, but stated only vaguely without leading the reader through the reasoning of the answer (i.e., "more flips leads closer to true mean" without saying anything about how fewer flips would help) or did not relate the idea to the numbers or concepts in the problem or had an error in part of the reasoning.
- 3 = Incomplete Answer: Uses other reasoning that could be helpful to show that the answer was more than a guess. This could include ideas of the law of large numbers that are very ambiguous.
- 2 = Incorrect Answer: Answers are based primarily on 50/50 chance or equal chance of heads and tails OR compare the RATIOs or percentage of Heads to Tails OR use law of large number reasoning to lead to an incorrect conclusion OR incorrect, ambiguous, or not helpful in another manner.
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as "don't know" rather than attempting an answer to the question

Item 15: (multiple-choice) A seventh-grade teacher in a school introduced his probability unit by giving his students a game involving one red die and one blue die. In the game, he would roll both dice once. The task for the students was to decide which of the following options would give them the best chance of winning this game. Which would you pick?

- 5 = You roll a 5 on the red die
- 3 = You roll a 5 on the red die, and a number other than a 6 on the blue die

- 2 = You roll a 5 on the red die, and a 6 on the blue die
- 1 = It doesn't matter. All of the previous three choices give the same chance of winning.

Item 16: (multiple-choice) You discovered a specially weighted coin that has a probability of 0.6 of landing on **TAILS** for any particular flip. If you flip this coin twice, what is the probability that it will land on **HEADS** both times?

- 5 = 0.16 (H*H)
- 4 = 0.36 (T*T)
- 3 = 0.25 (Fair H * Fair H)
- 2 = 0.8 (H+H)
- 1a = 0.4 (H)
- 1b = 0.6 (T)
- 0a = 0.5 (Fair H)
- 0b = 1 (Fair H + Fair H)
- 0c = 0 (impossible)
- 0d = 1.2 (T + T)

Item 17a: (multiple-choice) You see a woman carrying a baby. Which of the following is more likely?

- 5 = The woman is a doctor
- 3 = The woman is a doctor and a mother
- 1 = Both of these choices are equally likely

17b: (extended-response) Explain your reasoning in detail as if you were explaining to a 7th grade student.

- 5 = Desired Answer: The second choice adds a restriction OR the second choice is included in (a subset of) the first choice.
- 4 = Acceptable Answer: Explain how to compute the probabilities for the first two choices.
- 3 = Incomplete Answer: The idea of the answer implies one of the of the possibilities for a score of a 5 or 4, but the answer contains a significant conceptual error (other than not mentioning independence of items when mentioning multiplicative property) or is extremely vague OR only mentions the multiplicative property without explaining how or why it applies OR explanation without mathematical or statistical backing.
- 2 = Incorrect Answer: Context driven answers (for example: more doctors are men, or I have seen . . .) OR incorrect or incomplete in another manner
- 1 = Restatement of the question or multiple-choice part of the question without further explanation
- 0 = Participant did not answer the question or made a comment such as "don't know" rather than attempting an answer to the question.

VITA

Tamara Anthony Carter, 7777 South May Avenue, Oklahoma City, OK 73159-4444

EDUCATIONAL EXPERIENCE

Ph.D., **Texas A&M University**, Curriculum and Instruction with emphases in Mathematics Education and Educational Research, 2005.

M.S., **Texas A&M University**, Mathematics, 1999.

M.A., **Rice University**, Computational and Applied Mathematics, 1995.

Lifetime **Texas Secondary School Teaching Certificates** in mathematics and in computer information systems, 1994 and 1995, respectively.

B.A., **Rice University**, Computational and Applied Mathematics, 1994.

TEACHING EXPERIENCE

2004 – 2005 Mathematics Education Instructor, **Texas A&M University**.

2001 – 2003 Mathematics Lecturer, **University of North Texas.**

1999 – 2001 Mathematics Instructor, **Texas A&M University, Texas Woman's University,** then **University of North Texas**.

1995 – 1999 Mathematics Instructor then Associate Professor of Mathematics, **North Harris Montgomery Community College District**.

1995 – 1997 Mathematics Teacher, **Tomball High School**.

1993 & 1994 Mathematics Teacher, **Rice Summer School**.

SELECTED PUBLICATION

Carter, T. A., Tapia, R. A., & Papakonstantinou, A. (online). *An introduction to linear algebra: A curricular unit for pre-calculus students*. Retrieved October 25, 2004, from http://ceee.rice.edu/Books/LA/

SELECTED PAPERS SUBMITTED FOR PUBLICATION

Carter, T. A., & Capraro, R. M. (2005). *Stochastic misconceptions of pre-service K-8 teachers*. Paper presented at the annual meeting of the Association of Mathematics Teacher Educators, Dallas, TX.

Carter, T. A., Zientek, L. R., & Capraro, R. M. (2005). *Preservice Teachers' Understanding of Probability and Statistics*. Paper presented at the 32nd annual meeting of the Research Council on Mathematics Learning, Little Rock, AR.

Zientek, L. R., Carter, T. A., Taylor, J. M, & Capraro, R. M. (2005). *Prospective Teachers' Attitudes and Understandings of Statistical Concepts*. Paper presented at the at the 28th annual meeting of the Southwest Educational Research Association, New Orleans, LA.

SELECTED PRESENTATIONS

Kulm, G., Capraro, R. M., Capraro, M. M., Carter, T. A., Li, X., Sahin, A., et al. (2005, April). *How do students in the middle grades represent data?* Paper presented at the at the 83rd annual meeting of the National Council of Teachers of Mathematics, Anaheim, CA.

Capraro, R. M., Capraro, M. M., Harbaugh, A. P., Carter, T. A., Romero, C. T., & Naiser, E. (2005, April). *Using student achievement data to support teacher quality measures.* Paper presented at the research presession of the 83rd annual meeting of the National Council of Teachers of Mathematics, Anaheim, CA.