ABSTRACT

Title of Document: AN INVESTIGATION INTO STUDENT UNDERSTANDING OF STATISTICAL HYPOTHESIS TESTING

Toni Michelle Smith, Ph.D., 2008

Directed By: Dr. James T. Fey, Department of Curriculum and Instruction

In today's data driven world, the development of a statistically literate society is critical. As a result, many students are enrolling in university level introductory statistics courses and educators are promoting the development of strong understandings of the material taught in those courses. Statistical hypothesis testing, a powerful method of inferential statistics widely used in research, is taught in introductory courses. Though algorithmic in nature, statistical hypothesis testing is based on statistical theory. It is important that introductory students develop connected understandings of the algorithm, the concepts and logic that support it, and its uses.

This study explored the degree to which undergraduate, introductory statistics students develop desired understandings of the overall "big picture" of statistical hypothesis testing. In order to investigate student understanding a mixed methods approach was employed—both large scale quantitative and small scale qualitative data were collected. In the quantitative phase, a framework for assessing understanding of the conceptual and logical foundations of statistical hypothesis testing and its uses was created, a multiple-choice instrument with items representative of the framework was constructed, and data on student performance on this instrument were collected. Scores from a course exam that assessed student ability to use the algorithm to solve traditional statistical hypothesis testing problems were collected and compared with those from the multiple-choice instrument. In the qualitative phase, in order to gain more insight into student thinking, follow-up interviews were conducted with students who represent a range of performance patterns on the two quantitative assessments.

The data collected in this study indicated that introductory statistics students do not develop strong, connected understandings of the "big picture" of statistical hypothesis testing. Though they are able to perform the procedures, students do not have strong understandings of the concepts, logic, and uses of the method. A weak correlation between scores on the quantitative assessments indicated that procedural knowledge is not a predictor of overall understanding of statistical hypothesis testing. Analysis of quantitative and qualitative data indicated that students do not understand the role of indirect reasoning and inference in implementing and interpreting the results of a statistical hypothesis test.

AN INVESTIGATION INTO STUDENT UNDERSTANDING OF STATISTICAL HYPOTHESIS TESTING

By

Toni Michelle Smith

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Professor Emeritus James T. Fey, Chair
Associate Professor Patricia F. Campbell
Associate Professor Daniel I. Chazan
Professor Lewis Romagnano
Associate Professor Paul J. Smith

ACKNOWLEDGEMENTS

This dissertation would not have been possible if it had not been for the Mathematics Education program at the University of Maryland, College Park. The program offered a venue through which I was able to further develop my passion for mathematics education. Throughout, I was challenged to think hard about mathematics education and related fields. I learned how to design and conduct research studies to explore questions about the teaching and learning of mathematics and statistics. I was given the opportunity to participate in a variety of activities, both researching and teaching, that helped me to develop as a researcher and educator. I want to thank each member of the Mathematics Education faculty at the University of Maryland for these valuable experiences. I am truly grateful to have had this tremendous opportunity.

More specifically, I want to thank the director of this dissertation, Dr. Jim Fey, for all his help and guidance. His patience, support, and expert advice were truly appreciated throughout all phases of the study. He gave useful advice concerning the design of the study, provided valuable suggestions for the development of the multiple-choice assessment, and gave important feedback throughout the development of the chapters in this dissertation. Not only was Jim a valuable resource, but he was a wonderful cheerleader. He pushed my thinking about issues related to the teaching and learning of statistics and the study of those issues but gave me space to develop my own ideas. His passion for mathematics (and statistics) education is truly inspirational.

I would also like to thank the members of my dissertation committee. Each has played a role in the development of this dissertation and has been supportive of me throughout the process. As my academic advisor and a member of the committee, Dr.

Dan Chazan, offered useful advice for the development of this research project. He challenged me to think about what it means to “understand” and about how I might explore student understanding in the context of a follow-up interview. Dr. Patricia Campbell provided useful advice on the design of the study and helped me to think about data analysis within the context of my research questions. In addition, it was through my work with her on a large-scale quantitative research study that I was better able to do such work in my own dissertation. Dr. Lew Romagnano pushed my thinking about the concept of statistical hypothesis testing and what it means to understand this concept. And, as the multiple-choice assessment was being developed, Dr. Paul Smith provided valuable feedback on the items.

I would also like to thank the instructors of the introductory statistics course who allowed me to come into their classrooms to recruit participants and to administer the multiple-choice assessment. Additionally, I want to thank all of the participants in my study, particularly those who participated in the follow-up interviews. I appreciated their willingness to talk openly about their understanding of a difficult concept.

On a more personal note, I want to thank family and friends who offered me support and encouragement throughout the program and the dissertation process. Most importantly, I want to thank my parents, Gary and Kathy Smith, who, from an early age, instilled in me an appreciation for the value of education and the reward of hard work. Their belief in my capabilities and undying support were (and continue to be) truly appreciated. Finally, I would like to thank Aaron Schuetz who patiently endured countless conversations about the challenges and triumphs associated with this project. His support and encouragement were appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The field of statistics offers methods for using data to solve problems and inform decisions. In today's data driven society, these methods are of extreme importance. In particular, statistical methods provide a means for dealing with variability in data, so they play a large role in research and organized inquiry. Given a question of interest, statistical theories dictate the ways that sample data should be gathered, descriptive statistics should be used to describe and analyze the data, and inferential statistical methods should be used to draw conclusions about populations based on analysis of data samples. Due to the utility and power of statistical concepts and methods, individuals are in contact with statistics almost daily, either as consumers of information obtained from statistical study or as producers engaged in the research process.

In order to better prepare individuals to interpret and/or conduct research, courses in introductory statistics are offered in universities (Introductory Statistics) and high schools (Advancement Placement Statistics). Leaders in statistics and mathematics education have emphasized the need for introductory statistics students to understand not only the processes but also the logic and concepts that provide the foundation for statistical analysis, especially inferential methods. As a result, researchers are studying various aspects of the teaching and learning of statistical concepts. To advance this research agenda, work is needed on student understanding of one useful inferential method taught in introductory courses: statistical hypothesis testing.

This chapter introduces the study presented in this dissertation. The chapter is organized into three sections: (1) discussion of the background and rationale for study of understandings of statistical hypothesis testing; (2) presentation of conceptual frameworks and theoretical perspectives on which study of understanding may be grounded; and (3) presentation of research questions and methodology of the study.

Background and Rationale

The argument for studying student understanding of statistical hypothesis testing has its foundations in the general desire to develop a statistically literate and educated citizenry. Statistics plays an important role in modern society and, over time, educational leaders have begun to advocate for courses that aim to develop deep understandings of statistical concepts. When assessments of student performance call into question the degree to which key statistical understandings have been developed, research on student understanding can inform the design of more effective instruction. Such is the case for statistical hypothesis testing.

Statistical Analysis and Modern Society

Evidence of statistical reasoning can be found in everyday activity. Newspapers, magazines, journals, and television programs quote statistics in order to provide numerical information about stories. Statistical concepts are central to political polling reports, and, adoption of governmental policy is often influenced by statistics collected from census data (Franklin, Kader, Mewborn, Moreno, Peck, Perry, & Scheaffer, 2005). "Employment increasingly requires analytical, quantitative, and computing skills;" (Moore, 1997, p.124) and these skills are central to the practice of statistics.

The prevalence of statistics in modern society is largely due to its value in the inquiry process of research, where data are commonly used to answer questions of interest. Investigators observe the world around them, generate research questions, and collect data that will help them to answer their research questions. After analyzing the data, researchers are in a position to formulate conclusions in response to their research questions. Because statistics is the "branch of mathematics [that deals] with the collection, analysis, interpretation, and presentation of sets of data" (Lajoie, 1998, p. xii), statistical methods and concepts play a significant role in this process.

When data are used to answer a question of interest, it is important to recognize the variability that exists in that data. Statistical methods are designed to deal with this variability. In fact, Cobb and Moore (1997) claim that "the need for [statistics] arises from the *omnipresence of variability*" (Cobb & Moore, 1997, p. 801). Measurement variability, natural variability, induced variability, and sampling variability (Franklin, et al., 2005) are forms of variability that are associated with any data. The field of statistics provides researchers with methods of describing, measuring, and modeling this variability. These methods inform the ways in which researchers collect data, analyze data, and, ultimately, formulate conclusions based on the data (Reading & Shaughnessy, 2004).

In addition to providing a means of dealing with variability, a distinguishing feature of statistics is that it handles numbers in context. The motivation for statistical analysis is the context in which it is employed. Although the data are numerical, the numbers are connected to a context and "in data analysis, context provides meaning" (Cobb & Moore, 1997, p. 801). The design and ultimate interpretation of the results of studies that employ statistical analyses are highly dependent upon the context. The ability to deal with variability in context makes the field of statistics extremely valuable to modern society.

Statistical Understanding and the Role of Education

With the prevalence of statistical ideas in society, educators are beginning to outline those skills, understandings, and habits of mind that are important for individuals to function in a statistical society. Among those who have offered ideas, there is consensus that people should have an understanding of those statistical concepts and ideas that enable an individual to critically examine and interpret reports of statistical analysis. Individuals should have a basic knowledge of the ways that statistical inferences are made as well as an understanding of the ways in which descriptive statistics give information about a set of data (Gal, 2004).

Specific recommendations for K-12 statistical content and associated reasoning skills have been offered by leaders in the field (e.g. Franklin et. al, 2005; Lajoie, 1998; Scheaffer, Watkins & Landwehr, 1998). In addition, national organizations such as the National Council of Teachers of Mathematics (NCTM) have included statistical content in their recent reform efforts. In its *Principles and Standards for School Mathematics*, NCTM outlines objectives for the Data Analysis and Probability standard for students in grades K-12 (NCTM, 2000). School districts across the country are now including objectives that address statistical skills and understandings in their mathematics curricular frameworks.

The push for a statistically literate society is also being realized in upper levels of education, where courses are offered that provide instruction on more advanced statistical concepts and ideas. Universities offer Introductory Statistics courses and high schools now offer a course covering similar content in the form of an Advanced Placement Statistics course. The numbers of students taking these introductory courses has been rising as more and more students of increasingly diverse interests are enrolling in introductory and more advanced statistics courses (ASA, 2005). Recommendations for the teaching and learning of concepts in introductory courses have also been made by various leaders in the field (e.g. Garfield, 1995; Moore, 1997; Wild & Pfannkuch, 1999) and workgroups have been established to do the same.

In 1992, The Mathematical Association of America (MAA) published a document containing a series of recommendations concerning teaching and learning in introductory statistics courses. The group recommended that these courses should (1) emphasize statistical thinking; (2) include more data and concepts: less theory, fewer recipes; and (3) foster active learning (Cobb, 1992). Building on the work of the MAA Statistics Focus Group and funded by the American Statistical Association, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project stated that students who complete an introductory course should be "statistically educated." That is, "students should develop statistical literacy and the ability to think statistically" (ASA, 2005). This statement indicates a desire for students to develop both a deep understanding of statistical concepts and the reasoning and analytic skills necessary to the practice of statistics. Students should understand the concepts and ideas that support procedures used in the various methods and understand how these methods may be employed to answer questions of interest. Students should have an awareness of the role of context and variability in statistical analysis and should recognize the value of statistics in decision making (ASA, 2005).

Understanding of Inferential Statistical Methods: The Statistical Hypothesis Test

Given the value of inferential methods to the process of inquiry and given the recommendations for the development of statistically educated students in introductory classes, educational leaders have made recommendations for the teaching and learning of these methods as part of the introductory course. Implementation of formal inferential statistical methods in the process of inquiry involves the use of formulas, calculations, and reference to statistical tables. Though algorithmic in nature, these methods are supported by statistical theory. Both educational leaders and statisticians are recommending that students complete introductory statistics courses with deep understandings of these methods (e.g. ASA, 2005; Cobb, 1992; Garfield & Chance, 2000; Moore, 1997; Snee, 1999). Although procedures are important, a "grasp of the reasoning of inference is more important than how many individual procedures [statistics courses] touch on" (Moore, 1997; p. 127). According to Snee (1999), the development of statistical thinking skills is especially necessary in today's world where technology performs the procedures, leaving statisticians and consumers alike to critically examine the work presented in statistical investigations. These recommendations indicate that it is essential that introductory statistics students develop an understanding of the concepts and reasoning that support the procedures used in statistical analysis and inferential methods.

One inferential method that is extremely useful in research and that is included in the syllabi of introductory statistics courses is the *statistical hypothesis test*. Statistical hypothesis testing allows researchers to use information from one sample to place confidence in a given hypothesized description of an entire collection or group. Implementation of a statistical hypothesis test usually requires execution of a procedure in order to reach a conclusion. This procedure and ultimate conclusion are: (a) supported by theoretical ideas from statistics and probability and premises associated with logical proof; and (b) informed by the context in which the test is employed. As Moore (1997) and others (e.g. ASA, 2005) indicate, it is important that students complete introductory courses with a deep, connected understanding of statistical hypothesis testing that goes beyond the ability to perform the procedure. In addition, introductory students should also have the ability to consider the role of context in the design and interpretation of studies that use statistical hypothesis testing.

Though scarce in number, studies provide some evidence to suggest that individuals struggle in the development of a deep, connected understanding of statistical hypothesis testing and its use. Included in this literature are studies that focus on student errors in implementation of the procedure (e.g. Aquilonius, 2005; Evangelista & Hemenway, 2002; Hong & O'Neil, 1992; Link, 2002) and studies that study individuals' thinking and understanding of statistical hypothesis testing. These latter studies focus on: (1) the relationship of various components of hypothesis testing to the whole concept and process (Hong & O'Neil, 1992; Krauss & Wassner, 2002; Lane-Getaz, 2007; Lipson, Kokonis, & Francis, 2003; Mittag & Thompson, 2000; Wilkerson & Olson, 1997); and (2) the entire concept and process of hypothesis testing (Aquilonius, 2005; Liu, 2005).

Reports of student errors indicate that students struggle with almost every step in implementation of the procedure, indicating that they do not have a well developed understanding of statistical hypothesis tests (e.g. Link, 2002). This sentiment is echoed by the few studies of student understanding found in the literature (e.g. Aquilonius, 2005; Liu, 2005). The literature also includes theoretical commentaries that point to anecdotal evidence of student difficulties and suggest ways to confront those difficulties through instruction (e.g. Falk, 1986). But, even when instruction is specifically designed to decrease student errors when performing statistical hypothesis tests, Hong and O'Neil (1992) and Evangelista and Hemenway (2002) found that students continue to have difficulty. It seems that educators have not yet determined how to design instruction that will promote the development of deep, connected understanding of statistical hypothesis testing by introductory students.

A first step, then, is to construct a picture of student understandings of the "big picture" of statistical hypothesis testing that goes beyond a list of student errors. Though some researchers have begun this process (e.g. Aquilonius, 2005; Liu, 2005) and their studies have provided information useful in the construction of a picture of introductory student understandings of statistical hypothesis testing, these studies are scarce and limited in scope. More studies are needed to build on this work and further define the nature of student understanding. Therefore, research aimed at studying university-level, introductory statistics students' understandings of all aspects of statistical hypothesis testing would make a significant contribution to the field of statistics education as it strives to prepare students for daily and professional life in a statistical society.

Theoretical Perspective

In order to study introductory statistics students' understandings of statistical hypothesis testing, it is important to analyze the nature of statistical hypothesis testing itself and to define a framework useful for describing understandings of that concept. Analysis of the procedure and the conceptual foundation for that algorithm will inform the way in which a framework for understanding might be useful in the study of individuals' thinking about this important concept.

A Conceptual Analysis of Statistical Hypothesis Testing: The "Big Picture"

Hypothesis testing is used as researchers attempt to explain phenomena and to answer questions of interest. Once a question of interest has been identified, researchers make observations and begin to generate hypotheses representing possible answers to the question. In the interest of proving or disproving these hypotheses, researchers conduct studies in which they collect data for analysis. Based on their analysis, researchers make conclusions concerning the validity of their hypotheses. It is through coordination of various statistical concepts with probabilistic reasoning that statistical hypothesis testing becomes a powerful means of testing the validity of hypotheses using only information obtained from a sample of the population under study.

Proof and Hypothesis Testing Via a Sample

Premises of logical proof hold that in order to prove a statement about a group of cases, the statement must be shown to be true for *every single possible* case within that group. Unless exhaustive, the generation of supporting examples is not sufficient. As more examples are generated, one may become more confident that the premise is correct, but he/she cannot prove it unless it is demonstrated to be true for *all* cases. On the other hand, if one wishes to *disprove* a statement about all members of a group, he/she can do so by providing one counter-example. That is, one example can either (a) provide evidence for, but not a proof, of a given statement or (b) provide evidence against, and therefore disprove, a given statement.

It is often difficult to generate an exhaustive list of examples in order to prove a statement directly. Under these circumstances, it is possible to prove a statement using only one case, through indirect reasoning. Using indirect reasoning to prove a proposition *p*, one begins by assuming the negation of that proposition, ~*p*, and showing that such a supposition leads to a logical contradiction. Then the conclusion is that the original proposition, *p*, must be true. When the logic of indirect proof is applied to the challenge of establishing truth of a research hypothesis, one begins by stating the negation of that hypothesis (in statistical context, the *null hypothesis* or H_0), and looking for evidence showing that negation to be untenable. If such evidence is found, one can conclude that the original hypothesis (in statistical context, the *alternative hypothesis* or H_1), is true.

Although indirect reasoning provides a means of proof for the alternative hypothesis, it only does so if the evidence produced (in statistical context, a sample) is deemed impossible under the assumed null condition. Unfortunately, it is not always true that a given sample is absolutely impossible under the null hypothesis. Therefore, the researcher is not able to disprove the null hypothesis in favor of the alternative. He/she must draw an *inference* about the population based on the sample and his/her conclusion is accompanied by some degree of uncertainty.

In classical statistics, one draws an inference by determining the degree to which a sample presents characteristics that would be *unusual* if the null hypothesis did, indeed, describe the true nature of the population. That is, the researcher assumes that the null hypothesis describes the true nature of the population, and performs an analysis on the sample to determine whether that sample exhibits what he/she would *expect* (is a likely sample) or if the sample is *unusual* (unlikely). In the case where the sample is like what one would expect, the researcher has found support for the null and will *retain* (*fail to reject*) the null hypothesis. In the case where the sample is deemed unusual, the researcher has found evidence that the null is not the true descriptor of the population and will *reject* the null hypothesis in favor of the alternative hypothesis. In other words, the test essentially boils down to the following question and resulting action: *If the null hypothesis describes the true nature of the population, is the sample obtained considered to be unusual? If so, then reject the null hypothesis. If not, retain (fail to reject) the null hypothesis*. Note that, in either case, the researcher is not able to prove (or disprove) a given hypothesis. Using the analysis of a single sample, the researcher can only provide evidence for or against a given hypothesis.

In working through this line of reasoning, one must determine the criteria that will be used to decide whether or not the sample provides evidence against the null hypothesis. In order to do so, researchers rely on *test statistics*, which are numerical summaries of sample data characteristics. The researcher must decide which test statistic is appropriate for the analysis and this statistic is used to determine whether or not the sample is *unusual* under the null hypothesis. In making this decision, it is important to recognize the variability associated with sampling. Within the same population, a given sample of that population will look different from another sample taken from the same population. Analysis of those samples, then, will yield different results. The variability that exists among samples must be taken into consideration when attempting to decide whether one given sample is *unusual*.

Because the field of statistics provides a means of taking variability into consideration, it seems a good place to search for a way to consider variability in the decision to reject or retain the null hypothesis. In fact, statistics does offer such a concept: the sampling distribution of the statistic. The sampling distribution of the statistic connects probability to the sampling process and is thus useful in helping a researcher to determine whether or not a sample is unusual and to quantify the degree to which he/she is confident in the resulting decision to reject or retain hypotheses.

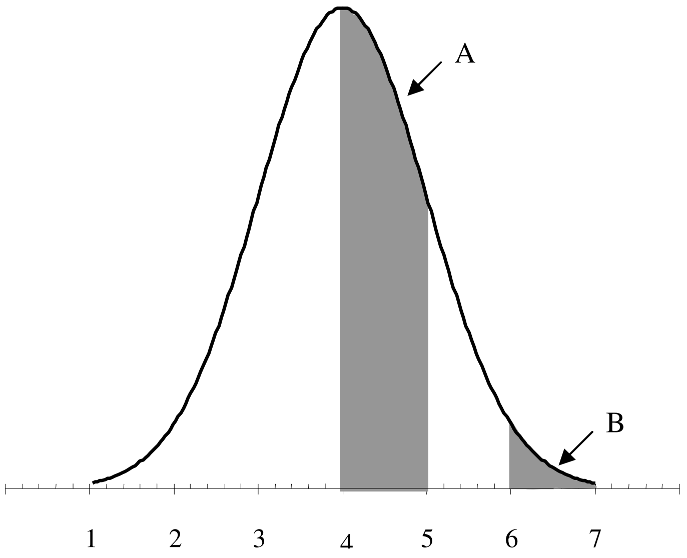
Unusual or Expected? Application of the Sampling Distribution

The *sampling distribution of the statistic* is exactly what the name implies. Given a sampling scheme producing samples of size *n*, a test statistic of interest (such as an average, proportion, or correlation), and all possible samples that may be obtained by randomly sampling the population, the sampling distribution of the statistic represents the distribution of the values of the statistic that are obtained from the samples. That is, for a given sample size, *n*, the sampling distribution gives the distribution (relative frequencies) of values of the statistic for all possible samples of *n* that result from random sampling. The resulting distribution can be modeled by a probability formula from which probabilities may be determined.

Consider the example shown in Figure 1.1 that is based on a sampling distribution of an average.

Figure 1.1

Sampling Distribution of an Average, Sample Size *n* = 50



Here, the probability that a randomly chosen sample of size 50 will have an average between 4 and 5 is the same as the area under the curve between 4 and 5 (shaded part A). The probability that a sample of size 50 will have a mean greater than 6 is the area under the curve from 6 to the right (shaded part B). Note that for a continuous variable, calculation of area is not possible for a single value of a statistic. Therefore, probabilities are typically determined for *intervals* of a continuously distributed statistic.

Because the sampling distribution of the statistic gives probabilities associated with obtaining certain values from a randomly chosen sample, it can be used to determine whether or not a sample is unusual (unlikely) in the population being modeled by the sampling distribution. Thus, sampling distributions are extremely valuable tools for use in hypothesis testing. An unusual sample would occur with a low probability. In attaching a probability value to the sample, the researcher is able to not only determine whether that sample is unusual but also to quantify the degree to which he/she is confident in his/her conclusion.

Unfortunately, the sampling distribution cannot be used unless the characteristic of interest in analysis of a sample is a measurable quantity, namely a statistic. For this reason, then, *statistical* hypothesis testing can only be used to test hypotheses that address some quantitative characteristic of the population, called the *population parameter*. The sample statistic can then be compared to a sampling distribution, connected to a probability, and a decision reached concerning whether or not it is unusual (unlikely).

Although it may seem that the sampling distribution is a factor that limits the usefulness of statistical hypothesis testing, it is actually one of the keys to the power of the method. Statistical hypothesis testing does <u>not</u> prove or establish a given hypothesis to be a true descriptor of the population. It does provide a means of quantifying the uncertainty associated with the inferences drawn from sample information. In order to better illustrate the way in which statistical hypothesis testing gives the researcher the ability to quantify his/her uncertainty, the entire process will now be outlined.

Statistical Hypothesis Testing: Using Probability to Make Inferences

Given a research question that concerns a large population, a researcher must identify an appropriate population parameter to consider which would answer the question. Then, the researcher establishes both the null and alternative hypotheses to describe the population parameter under question. The researcher generally constructs these hypotheses so that the alternative aligns with a statement about the population which the researcher wishes to prove, while the null hypothesis is that which is contradictory to that statement. The researcher's goal is to provide evidence that the null hypothesis is not correct. To do so, the researcher assumes the null to be a true description of the population in hopes that analysis of a sample will contradict that notion.

The researcher obtains a randomly chosen sample and calculates the sample statistic corresponding to the hypothesized population parameter. Given the sample size, there exists a sampling distribution of that statistic associated with the population that has the parameter value stated in the null hypothesis. Using the appropriate sampling distribution, the researcher can determine the probability of obtaining his/her sample statistic or more extreme (more unlikely – more distant from the hypothesized value) under the condition that the null hypothesis is the true descriptor of the population. If the probability is small, the sample is deemed unusual under the null hypothesis. Using indirect reasoning, then, the null hypothesis is rejected and more confidence is placed in the competing, contradictory alternative hypothesis. If the probability is large, then the sample is considered to be typical of what is expected under the null hypothesis and the null hypothesis is retained as there is not enough evidence to believe otherwise.

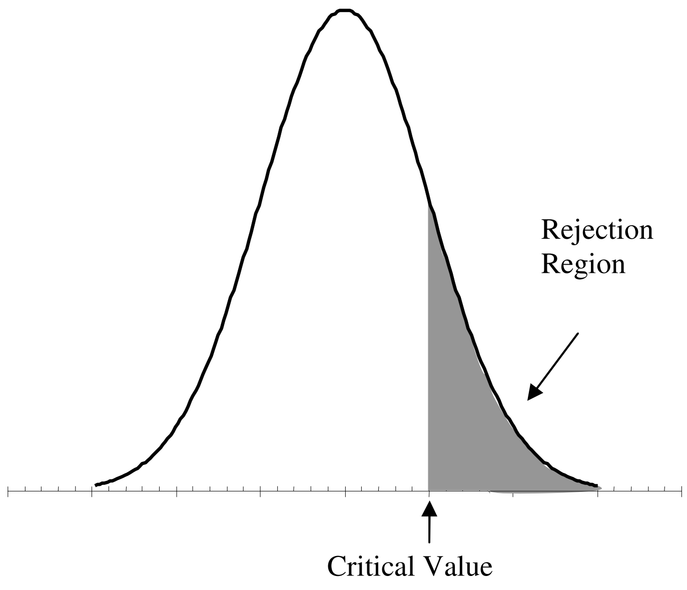
In order to decide if the probability is small or large, the researcher must determine a decision rule that designates a "cut point." That is, he/she must decide what probabilistic value is considered small and would lead to a rejection of the null hypothesis. For example, a decision rule might be: If the probability associated with obtaining a sample whose sample statistic is x or more extreme is smaller than 0.05, the null hypothesis will be rejected; if not, the null hypothesis will be retained (or will not be rejected). A decision to reject the null hypothesis indicates that under the condition that the null hypothesis is true, the probability of obtaining the observed result (or more extreme) is only 0.05 (Shaver, 1993). The value, 0.05, is called the level of significance and is commonly denoted by the Greek letter, α.

Not only does the decision rule guide the formulation of a conclusion, but the decision rule also specifies the degree of uncertainty associated with making that conclusion. If the decision rule uses a significance level of $\alpha = 0.05$, then if the null hypothesis described the true nature of things and the researcher repeated his/her experiment over and over, he/she would obtain a sample that would lead him/her to incorrectly reject the null hypothesis five percent of the time. Thus, if the null hypothesis were a true description of the population, and the researcher obtains a sample that has a probability of occurring 0.05 or less, then he/she would *incorrectly* reject the null (Shaver, 1993). This probability is commonly referred to as the Type I error (see, for example, Hubbard & Bayarri, 2003). It is important to note that α is <u>not</u> the probability that the null hypothesis is true. It is the long run probability of rejecting the null hypothesis given that it is, indeed, true.

Some key terms that are associated with the decision rule are the terms *critical value* and *p-value*. Some statisticians and textbooks make a distinction between these two ideas in regards to performance of a hypothesis test. The *critical value(s)* refers to the value of the sample statistic that marks the “cut point(s)” used in the decision rule. That is, once a value for α has been specified, a *rejection region* is outlined on the sampling distribution. Values more extreme than the critical value lie in the tails of the distribution (smaller probabilities are attached to values that appear in the tails). In the case where the null and alternative hypotheses contradict each other in one direction (for example the null may claim the mean to be 4, while the alternative claims it to be larger than 4), the extreme values lie to the right of the critical value. Thus, the rejection region is also to the right of the critical value. The critical value and associated rejection region are illustrated in Figure 1.2.

Figure 1.2

Critical Value and Associated Rejection Region



In the case where the null and alternative hypotheses contradict each other in two directions (for example, the null may claim the mean to be 4, while the alternative claims that it is simply not equal to 4), there are two critical values. One critical value is greater than 4 resulting in extreme values and associated rejection region to the right of this critical value. The other critical value is less than 4 with extreme values and associated rejection region to the left of this critical value. The critical values mark the boundaries for rejection regions. Given the critical value(s), one must only compare the sample statistic with the critical value(s) in making a conclusion. If the sample lies in a rejection region, then the null hypothesis is rejected in favor of the alternative. If not, then the null hypothesis is retained. Huberty (1993) refers to this approach as the *fixed-α-approach* and credits this method to Jerzy Neyman and Egon Pearson.

The *p-value* indicates the actual probability that one would obtain a sample statistic as extreme, or more extreme, than the observed value, given that the null hypothesis holds. Once a sample statistic is obtained, the probability of obtaining that value or more extreme given the null hypothesis is determined and is labeled the *p-value*. If this value is small, the null hypothesis is rejected. Huberty (1993) refers to this approach as the *p-value* approach or as *significance testing* and credits it to R.A. Fisher.

The two approaches to statistical hypothesis testing developed by Fisher and Neyman and Pearson are based on two different philosophies of hypothesis testing, and were the subject of debate between the men in the mid-1900s (Hubbard & Bayarri, 2003). Although many statisticians consider the different approaches to hypothesis testing to be a “non-issue” (Huberty, 1993) and tend to merge them in practice (Batanero, 2000; Hubbard & Bayarri, 2003), it is important to address these differences in this conceptual analysis.

In his approach to statistical hypothesis testing, Fisher did not include the formulation of an alternative hypothesis, nor did he explicitly consider the cut point, $\alpha = 0.05$, to represent long run error rate. In his approach, the data are evaluated to determine the degree to which they provide evidence against the null hypothesis. Assuming the null hypothesis, sampling distributions are used to determine the degree to which the sample deviates from the mean of the sampling distribution associated with the null. The *p*-value described above provides a measure of that evidence. Smaller *p*-values indicate stronger evidence against the null. If a test yields a small *p*-value (and 0.05 was commonly accepted as the cut point), then either the data are very rare, or the theory (the null hypothesis) is not true. Fisher's method is one of *inductive inference* in that it provides a means for drawing inferences about a population based on information from a sample (Hubbard & Bayarri, 2003).

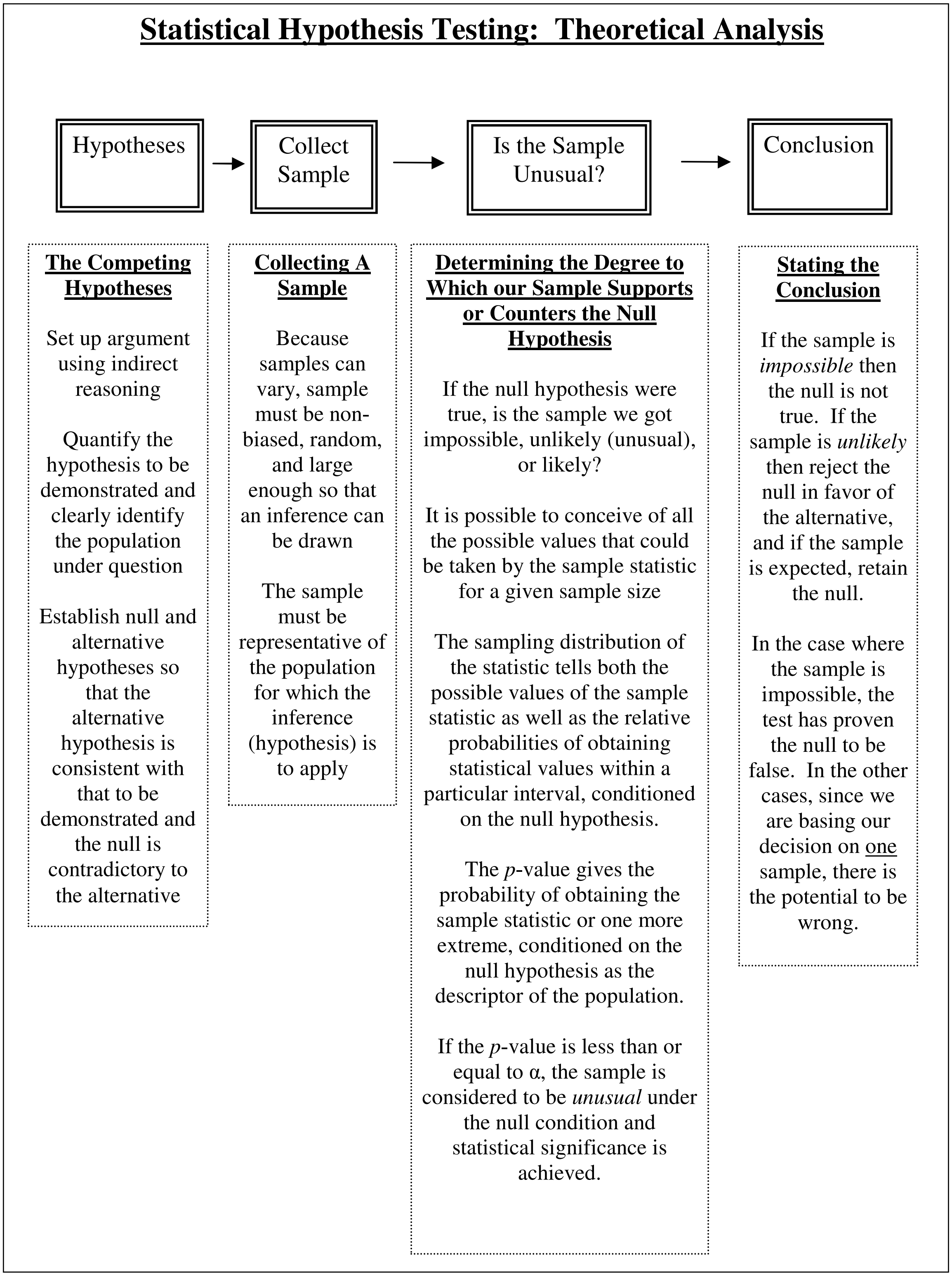
In contrast to Fisher's approach, the establishment of an alternative hypothesis is critical in the Neyman-Pearson approach. Neyman-Pearson conceptualized their approach as a test that places two hypotheses (the null and the alternative) in opposition. The ultimate goal is to make a decision between these two hypotheses in light of the data collected. In so doing, probabilities of two forms of potential error are introduced: Type I and Type II. Type I error (α) is the probability of incorrectly rejecting the null hypothesis and Type II (β) of incorrectly accepting the null hypothesis. In addition, the Neyman-Pearson approach introduces the concept of power (1 – β), which is the probability of correctly rejecting the null (Hubbard & Bayarri, 2003).

The Neyman-Pearson method uses pre-set, context dependent values for α and β to identify critical values and rejection regions to make that decision. "This is in sharp contrast to the data-based *p*-value, which is a *random variable* whose distribution is uniform over the interval [0, 1] under the null hypothesis" (Hubbard & Bayarri, 2003, p. 173). This approach does not use *inductive inference*. Rather, it is what Neyman terms *inductive behavior* (as cited in Hubbard & Bayarri, 2003). Decisions about how to proceed (or behave) are made using a limited amount of information. Sample information is used to decide whether to act as if (a) the null hypothesis were a true description of the population or (b) the alternative hypothesis is a true description of the population. Thus, the behavior is guided by a degree of inductive reasoning. However, the decision to act according to the null hypothesis or the alternative hypothesis is dictated by a set of rules that rely on statistics and probability. Therefore, the approach is deductive in nature, as it "argues from the general to the particular" using established rules "for choosing between two alternative courses of action, accepting or rejecting the null hypothesis" (Hubbard & Bayarri, 2003, p. 273). In the long run, operation by these rules should result in decisions that are more often right than wrong (Hubbard & Bayarri, 2003).

Though these differences are important, the overall ideas remain the same and, as mentioned above, are commonly merged in practice and instruction (Batanero, 2000; Hubbard & Bayarri, 2003). Hypotheses are established that address a quantitative feature of the population under study and a sample is collected for analysis. Using the sampling distribution of the statistic and probability theory, a decision is made concerning the degree to which the sample is unusual conditioned on the null, and a conclusion about the validity of the hypotheses is stated. Although it is not possible to prove a claim through inferential reasoning, statistical hypothesis testing relies upon the sampling distribution and probability to provide the researcher with a means to quantify the degree to which he/she is confident in his/her conclusion. The relationship between the various components of statistical hypothesis testing with each other and with testing of general hypotheses is illustrated in Figure 1.3.

Figure 1.3

Theoretical Analysis, Statistical Hypothesis Testing



Statistical Hypothesis Testing: Theoretical Analysis

The Competing Hypotheses

Set up argument using indirect reasoning

Quantify the hypothesis to be demonstrated and clearly identify the population under question

Establish null and alternative hypotheses so that the alternative hypothesis is consistent with that to be demonstrated and the null is contradictory to the alternative

Collecting A Sample

Because samples can vary, sample must be non-biased, random, and large enough so that an inference can be drawn

The sample must be representative of the population for which the inference (hypothesis) is to apply

Determining the Degree to Which our Sample Supports or Counters the Null Hypothesis

If the null hypothesis were true, is the sample we got impossible, unlikely (unusual), or likely?

It is possible to conceive of all the possible values that could be taken by the sample statistic for a given sample size

The sampling distribution of the statistic tells both the possible values of the sample statistic as well as the relative probabilities of obtaining statistical values within a particular interval, conditioned on the null hypothesis.

The *p*-value gives the probability of obtaining the sample statistic or one more extreme, conditioned on the null hypothesis as the descriptor of the population.

If the *p*-value is less than or equal to α, the sample is considered to be *unusual* under the null condition and statistical significance is achieved.

Stating the Conclusion

If the sample is *impossible* then the null is not true. If the sample is *unlikely* then reject the null in favor of the alternative, and if the sample is expected, retain the null.

In the case where the sample is impossible, the test has proven the null to be false. In the other cases, since we are basing our decision on one sample, there is the potential to be wrong.

The analysis of statistical hypothesis testing presented above illustrates the nature of statistical hypothesis testing as a method of inquiry that is highly reliant on indirect and probabilistic reasoning. It is important to note that statistical hypothesis testing, as implemented, is a procedure. The procedure requires calculations of sample statistics and use of tables (or more recently computing software) to determine probabilities. Even the conclusion is reached as a result of following a rule. When executing a statistical hypothesis test, one only needs to follow the prescribed steps in the procedure. The steps for the *p-value* and *fixed-α-approaches* are

Significance Testing	*Hypothesis Testing*
1. State the null hypothesis	1. State the null and alternative hypothesis
2. Specify test statistic (T) and referent distribution	2. Specify test statistic (T) and referent distribution
3. Collect data and calculate value of T	3. Specify α value and determine rejection region (R)
4. Determine p-value	4. Collect data and calculate value of T
5. Reject null hypothesis if p-value is small; otherwise retain	5. Reject null hypothesis is favor of alternative if T value is in the rejection region; otherwise retain null

(Huberty, 1993, p. 318)

In practice, however, these two approaches are generally merged. The decision to reject the null hypothesis based on a small *p*-value is equivalent to choosing a value for α in advance.

The Role of Context

The conceptual analysis of statistical hypothesis testing performed thus far has focused on the theoretical foundation of the steps involved in conducting a statistical hypothesis test. The role of context, while alluded to, has not, to this point, been explicitly described. Without context, this method would be meaningless. Therefore, it is important to discuss how context and theory interact so that the algorithm has meaning beyond its theoretical foundation.

Prior to conducting a statistical hypothesis test, an essential first step is to recognize whether it is possible and/or of value to use statistical hypothesis testing to answer the research question of interest. A researcher must have generated a research question as well as a hypothesized answer to that question. The researcher must then determine whether it is possible to answer the question in terms of a parameter of the population. Some questions cannot be answered by a quantifiable measure and thus, statistical hypothesis testing cannot be employed as no statistical inference is possible. For those that can, it is important to choose an appropriate measure so that the test can be used to answer the question. In some cases a mean is useful to answer the question and, in others, a proportion. It depends upon the context and the nature of the available data. If the question and hypotheses can be quantified and if it is necessary to test the hypothesis using only analysis of the probability associated with a sample (it is impossible to collect information for every member of the population in question), then statistical hypothesis testing is applicable. Overall, the decision to use statistical hypothesis testing is dependent upon a set of criteria that are all context bound.

In addition, to being quantifiable, the hypotheses under consideration in the hypothesis test must be "reasonable." Indirect reasoning follows the structure of the *modus tollens* argument: if p then q; not q; then not p. Modification of this format for statistical hypothesis testing is structured as follows: if p then probably not q; q; then probably not p. These logical symbols translate to: if the null hypothesis holds then particular samples would be unlikely; using random sampling, one of those samples is produced; then there is less confidence in the null hypothesis. However, as Cortina and Dunlap (1997) point out, the argument does not hold if the truth of the antecedent, *p*, and consequent *q* are not related, or are negatively related. Cortina and Dunlap (1997) give the following example: "If a person is an American, then that person is probably not a member of Congress; This person is a member of Congress; This person is probably not an American" (p. 165). The example illustrates the need for a reasonable relationship between the antecedent (in this case, "If the person is an American") and the consequent ("that person is probably not a member of Congress"). The consequent holds in many different statements for the antecedent. Therefore, researchers must be careful in designing their hypotheses so that there is a reasonable connection between the antecedent and consequent in the null hypothesis (Cortina & Dunlap, 1997).

Another point at which context must be considered is in the statement of a decision rule. As was described above, the level of significance indicates the degree to which the researcher is comfortable in making what Neyman and Pearson term a Type I error. Specifically the level of significance (Type I error) is the probability that, in the case where the null hypothesis is true, the researcher will incorrectly reject it in the long-run (Hubbard & Bayarri, 2003). The level at which this potential for error is set, again, depends upon the context. The researcher must decide the cost of making such an error and balance that cost with the actual statement of the decision rule. In addition, if the researcher is interested in using the study to make a strong argument to a group of people, the researcher must consider the standard set by that larger community. Within researching communities, the standard level of significance is commonly set at $\alpha = 0.05$ or at $\alpha = 0.01$ (Frick, 1996).

In addition to the probability associated with making a Type I error, another error one could make is to retain the null when it is, indeed, false. This is a Type II error. Type II error is often reduced by using a larger sample in the study. Therefore, decisions concerning sample size are context dependent and are influenced by the degree to which various levels of Type II error are acceptable (Hubbard & Bayarri, 2003).

Consideration of context is ultimately important in the formulation of the conclusion statement and subsequent decision making that is based on that conclusion. Here, knowledge of the conceptual foundation for the test itself interacts with the context in order to best interpret the results of the study. The degree to which a given study has potential for error may impact the decisions one makes based on the results of the study. Understanding that the test does not constitute a proof has implications for the degree to which an individual should make decisions that are influenced by the results of the study. Therefore, sample size and potential for error should be weighed against the cost of making that error when making the decision to act based on the results. The cost of making an error is, of course, context dependent and should be taken into consideration when deciding to what degree the results will influence future decisions.

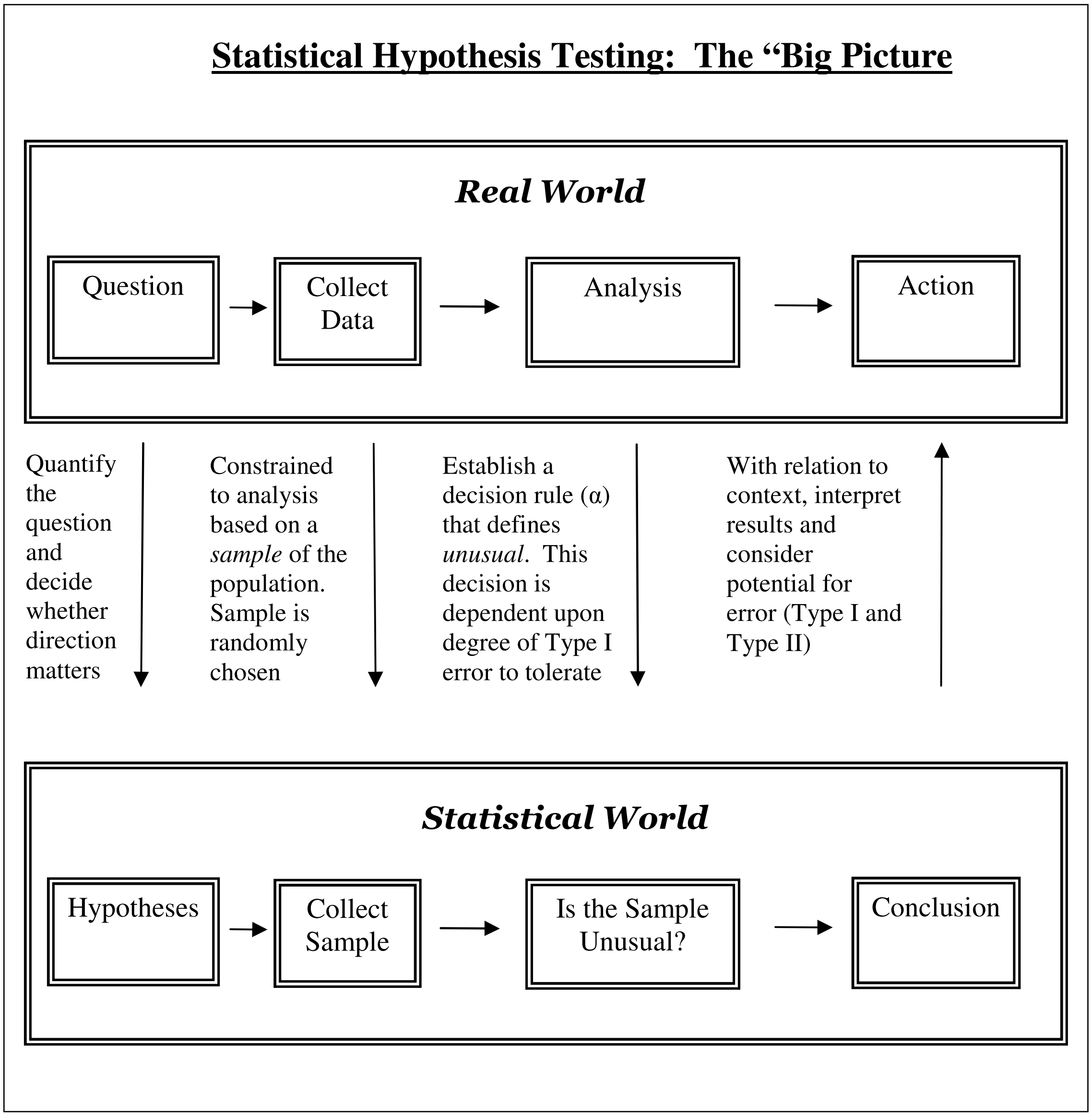
It is important to note that although the potential for a Type I or Type II error is present in all statistical hypothesis tests, it is not true that both errors can/will be committed. Each is conditioned on the property that the null hypothesis is or is not true. Therefore, "alpha is not the probability of making a Type I error. It is what the probability of making a Type I error would be if the null were true" (Cortina & Dunlap, 1997, pp. 166-167). The same could be said about the probability of a Type II error but with respect to the condition that the null hypothesis were not true.

In addition to the potential for Type I and Type II error, other aspects of a given study should be taken into consideration when considering the impact of the results on practical application. Statistical significance does not always imply scientific (or practical) significance. It is, indeed, possible to obtain significance in cases where the actual difference is very small and may not warrant action. Robinson and Levin (1997) give an example of a *t*-test comparison of the effects of a GRE preparation program. The test showed a significant difference in the mean scores of individuals who participated in the program versus those that did not, with a *p*-value of 0.043. However, the actual difference in scores for the samples of size 800 is 13 points. This difference is small considering that scores on this test range from 400 to 1600. Robinson and Levin (1997) warn that the money needed to participate in the program may not be worth a 13 point improvement. In this case, the effect size is very small and though participation might improve scores, the cost may not be worth it (Robinson & Levin, 1997).

As illustrated in this section, the role of context in the design and interpretation of studies that use statistical hypothesis testing is significant. A research question and accompanying hypotheses are established in the real world. In order to use statistical hypothesis testing, the hypotheses must be transformed. Once transformed, statistical analysis is performed and a conclusion reached based on that analysis. Ultimately, though, the interpretation of the conclusion must consider the way in which that conclusion is useful in the real world. Building on the diagram presented in the previous section (Figure 1.3), the diagram pictured in Figure 1.4 illustrates the connection of context to the process of statistical hypothesis testing in the overall framework of the "big picture" of statistical hypothesis testing.

Figure 1.4

Conceptual Analysis, Statistical Hypothesis Testing



Statistical Hypothesis Testing: The “Big Picture

Quantify the question and decide whether direction matters

Constrained to analysis based on a *sample* of the population. Sample is randomly chosen

Establish a decision rule (α) that defines *unusual*. This decision is dependent upon degree of Type I error to tolerate

With relation to context, interpret results and consider potential for error (Type I and Type II)

A conceptual analysis of the “big picture” of statistical hypothesis testing has now been presented. This analysis has highlighted the algorithmic and conceptual nature of the method. In addition, the analysis illustrated the ways in which the method is applied in context. A deep, connected understanding of statistical hypothesis testing (an understanding of the “big picture”) must include knowledge that supports use of the procedure, knowledge of the underlying reasoning and associated components, and an ability to consider context in application of statistical hypothesis tests in real world settings.

Describing Understanding of Statistical Hypothesis Testing

What is meant by the term understanding? More specifically, what does it mean to understand the concept of statistical hypothesis testing? The former question has plagued both cognitive psychologists and educators for a long time, while the latter is of particular interest to this study. Leaders in cognitive psychology and education have spent a considerable amount of time studying human cognition in the attempt to formulate definitions for "knowledge" and "understanding" and to identify the ways in which an individual comes to understand concepts and ideas. The result of this line of inquiry is the development of a body of literature filled with a variety of definitions of understanding and theories of cognitive activity that support learning and the construction of knowledge. Some of these theories seem particularly relevant to describing the nature of understandings that students have of statistical hypothesis testing.

Research on student understanding of statistical hypothesis testing has provided evidence that different kinds of understandings can emerge. These findings support the notion that an ability to execute the procedure for solving a well defined statistical hypothesis testing problem does not necessarily imply that the student knows why the steps are necessary and/or how this process can be applied in ill-defined, real world contexts (e.g. Aquilonius, 2005; Liu, 2005).

These distinctions are also evident in the conceptual analysis. Statistical hypothesis testing is largely a method of inquiry that relies on probabilistic and logical reasoning to connect various concepts within statistics. Like all inferential statistical methods, use of this method to answer real world questions requires interplay between the theory and the context in which it is employed. Therefore, it is important that analysis of an individual's understanding of statistical hypothesis testing addresses his/her ability to coordinate various ideas and ways of reasoning, *as well as* his/her ability to apply that knowledge in real life situations in a way that considers context. A perspective for describing understandings of statistical hypothesis testing should draw from theories of knowledge and/or understanding that address not just an individual's ability to execute the steps of the procedure, but also whether that individual knows *why* the steps are important and *how* the theory and procedure are applied in context. Taken together, several theories of thinking and learning can provide such a perspective.

In the interest of describing human thinking, cognitive psychologists have offered various theories that describe the nature of the internal structures or mechanisms that support the acquisition, organization, storage, and retrieval of information (Sternberg, 1999). Within the context of mathematics, Hiebert and Carpenter (1992) draw on this notion of internal structures and claim that as individuals are exposed to new mathematical concepts, they create internal representations of these concepts and, over time, form connections between the representations. The formation of these connections leads to the development of internal networks (metaphorically likened to vertical hierarchies or webs) which are linked to the development of understanding. "A concept is understood if it is part of an internal network. More specifically, the mathematics is understood if its mental representation is part of a network of representations" (Hiebert & Carpenter, 1992, p. 67). Internal networks are dynamic and are constantly being reorganized and reconfigured by individuals. As individuals construct more coherent and rich webs of connections and relationships among representations for ideas and concepts, they increase their level of understanding (Hiebert & Carpenter, 1992).

Using this framework for understanding, Hiebert and colleagues outline two categories of mathematical knowledge: conceptual and procedural. Conceptual knowledge is "knowledge that is rich in relationships. It can be thought of as a connected web of knowledge, a network in which the linking relationships are as prominent as the discrete pieces of information" (Hiebert & Lefevre, 1986, p. 3). A unit of knowledge is only conceptual if it is linked to other knowledge. Development of conceptual knowledge occurs when relationships are formed either between two pieces of existing knowledge or between a new and existing piece of knowledge. The relationships that are formed exist on different levels of abstraction. Relationships that go beyond superficial features and that are less tied to context are more abstract (Hiebert & Lefevre, 1986). As individuals construct relationships at a higher degree of abstraction, they create more coherent, connected internal networks and, thus, their level of understanding increases.

Procedural knowledge encompasses knowledge of (1) the formal system of mathematics and (2) knowledge of the algorithms or rules used in solving problems and doing other mathematical exercises. Such knowledge does not address meaning, only surface features associated with the form of mathematics, such as symbolic manipulation. Relationships between units of procedural knowledge are hierarchically arranged such that representations for smaller procedures are held under the umbrella of a representation for a larger, more encompassing procedure. The form of the networks associated with procedural knowledge is, therefore, different than that for conceptual knowledge.

Procedural knowledge is connected by hierarchical relationships whereas conceptual knowledge is connected by a host of varied relationships (Hiebert & Lefevre, 1986).

Hiebert and colleagues (1986, 1992) suggest that deep relationships exist between procedural and conceptual knowledge. For example, conceptual knowledge can give meaning to the symbols associated with procedural knowledge while, on the other hand, the use of symbols can make it less cumbersome to deal with conceptual ideas, allowing the user to further develop conceptual knowledge. Hiebert and colleagues maintain, therefore, that it is difficult to dichotomously classify knowledge as procedural or conceptual. Some units of knowledge fall into both categories. Regardless of the classification, though, it is important that learning goals address both forms of knowledge (Hiebert & Lefevre, 1986; Hiebert & Carpenter, 1992).

The definitions for procedural and conceptual knowledge offered by Hiebert and colleagues (1986; 1992) provide mathematics educators with a theoretical perspective with which to talk about understanding of mathematics, however some researchers contend that their definitions are not complete. Star (2005) argues that Hiebert and Lefevre's (1986) definition for procedural knowledge points to rote knowledge of algorithmic procedures, but does not encompass knowledge of and ability to apply more general procedures, such as heuristics. In addition, "flexible" knowledge of procedures and how/when to use them when solving novel problems is not accounted for in Hiebert and Lefevre's (1986) definition. Star (2005) also challenges Hiebert and Lefevre's (1986) definition of conceptual knowledge, claiming that their definition only accounts for conceptual knowledge that is richly connected. Star (2005) argues that the term *concept* itself (and he gives the example of the concept of *dog*) does not necessarily imply connections. Star (2005) contends, therefore, it is not sufficient to define conceptual knowledge as rich in relationships and procedural knowledge as rote knowledge of algorithms for which connections exist only among the steps of the algorithms. Procedural knowledge can also be rich in connections, and conceptual knowledge may/may not (Star, 2005).

Star's (2005) challenge of Hiebert and Lefevre's (1986) definitions suggests a distinction between knowledge type and knowledge quality. In response, Baroody, Feil, and Johnson (2007) propose that connectedness between and within procedural and conceptual knowledge types are important (and necessary) components of deep knowledge and understanding. In addition, Baroody, Feil, and Johnson (2007) argue that quality of knowledge should not only be measured by degree of connectedness. Ability to apply knowledge in a variety of situations: well or ill defined situations, real world or abstract situations, etc. is also important. According to Baroody, Feil, and Johnson (2007), depth of knowledge and understanding increases only when both (1) connections are formed within and between conceptual knowledge **and** (2) ability to use that knowledge adaptively and flexibly in a variety of settings increases (Baroody, Feil, & Johnson, 2007).

The definitions for knowledge types and their role in the development of understanding presented initially by Hiebert and colleagues (1986; 1992) with later refinements offered by Star (2005) and Baroody, Feil, and Johnson (2007) seem relevant to statistical understanding in general, and to understanding of statistical hypothesis testing more specifically. In addition to an understanding of the procedures and underlying theoretical foundation, statistical reasoning and thinking involves consideration of the context in which it is employed. The design and ultimate interpretation of studies that use statistical analysis should be developed with regard to the context in which they are employed.

In fact, several statistical educators have attempted to categorize these abilities. In their description of statistical thinking, Wild and Pfannkuch (1999) identified dimensions of statistical thinking that are important in using statistics to answer questions of interest. For them, statistical thinking involves an ability to engage in investigation (state a problem, make a plan, collect data, analyze the data, draw a conclusion, and continue in that cycle) and engage in interrogation (generate potential models or explanations, seek information, interpret that information, criticize the information, and judge whether to believe or use the information). In addition statistical thinking involves specific dispositions (e.g. skepticism, curiosity, being logical) and ways of thinking (e.g. recognition of the need for data, ability to transform data into measures, consideration of variation, reasoning with statistical models, and integrating the statistics with the context). These ways of thinking are important to the practice of statistics (Wild & Pfannkuch, 1999).

In a 2002 publication of the *Journal of Statistics Education* several articles were written to define various habits of mind associated with statistical practice. In this issue, researchers outlined three categories: statistical literacy, statistical reasoning, and statistical thinking. Rumsey (2002) describes statistical literacy as encompassing a basic knowledge of statistical concepts and terminology as well as an ability to read and interpret statistical analysis. Garfield (2002) describes statistical reasoning as an ability to use understandings of statistical concepts in order to draw conclusion and make inferences. She claims that "much of statistical reasoning combines ideas about data and chance, which leads to making inferences and interpreting statistical results. Underlying this reasoning is a conceptual understanding of important ideas…" (Garfield, 2002, section 1, ¶ 5) Finally, Chance (2002) describes statistical thinking as "what a statistician does" (section 2, ¶ 14). Statistical thinking involves seeing the big picture of raising questions, finding ways to answer them, understanding the role of variability and context in the statement of a conclusion, and continuing to engage in the inquiry process (Chance, 2002).

It seems there is some overlap between the constructs of statistical literacy, reasoning, and thinking and these ideas overlap with the various dimensions identified by Wild and Pfannkuch (1999). delMas (2002) gives possible frameworks for describing the overlaps.

One perspective which really seems to incorporate a good deal of the ideas presented above, and that seems an appropriate lens from which to think about student understanding of statistical hypothesis testing is that offered by the Mathematics Learning Study Committee of the National Research Council in the book *Adding It Up*, which was edited by Kilpatrick, Swafford and Findell (2001). In this book, the committee identifies five "strands" of mathematical proficiency:

- *conceptual understanding* – comprehension of mathematical concepts, operations and relations
- *procedural fluency* – skill in carrying out procedures flexibly, accurately, efficiently, and appropriately
- *strategic competence* – ability to formulate, represent, and solve mathematical problems
- *adaptive reasoning* – capacity for logical thought, reflection, explanation, and justification
- *productive disposition* – habitual inclination to see mathematics as sensible, useful, and worthwhile, coupled with a belief in diligence and one's own efficacy

(Kilpatrick, Swafford, & Findell, 2001, p. 5)

To illustrate the importance and interconnectedness of each strand to mathematical proficiency, the committee presents a picture of a rope in which each of the strands outlined above is intertwined (Kilpatrick, et al., 2001). This model points to the importance of the conceptual and procedural knowledge and degrees of understanding discussed by Hiebert and Lefevre (1986); Hiebert and Carpenter (1992); Star (2005); and Baroody, Feil, and Johnson (2007) while acknowledging the importance of an ability to use that knowledge flexibly in solving (and posing) problems that may or may not be situated within a greater context.

Although writers of *Adding It Up* (Kilpatrick, et al., 2001) were addressing what it means to be proficient in mathematics, this definition seems useful to describing how students understand or are proficient in statistics. Furthermore, given the conceptual analysis of statistical hypothesis testing developed in the previous section, this model seems an appropriate lens for analyzing and describing student understanding of the procedures, the logic, the various statistical and probabilistic components, and the role of context as well as how statistical hypothesis testing can or can not be used to answer research questions. Individuals must see the value in statistical hypothesis testing for answering questions of interest (*productive disposition*). They should recognize that the power of statistical hypothesis testing lies in its ability draw inferences about a population based on information from a sample and the way in which this is done (*conceptual understanding*, *adaptive reasoning*, some *productive disposition*, and some *procedural fluency*). In order to use statistical hypothesis testing, one must formulate the research question appropriately and employ the procedure (*strategic competence* and *procedural fluency*). In order draw a conclusion and interpret that conclusion with respect to the context, one must understand the concepts and logic involved (*procedural fluency*, *conceptual understanding*, *adaptive reasoning*, and some *strategic competence*). These understandings must be coordinated so that one informs the other, just as the rope illustrates.

A Study of Student Understandings of Statistical Hypothesis Testing

Armed with a conceptual analysis of statistical hypothesis testing and a perspective from which to describe understanding, it is now possible to outline a study of introductory statistics students' understandings of statistical hypothesis testing which will contribute to the construction of a more complete picture of student understanding of the overall concept.

Although scarce, studies of student understanding of statistical hypothesis testing have indicated that students struggle with virtually every step when implementing the procedure and students do not have deep understandings of statistical hypothesis testing. These results indicate that students do not develop connected, deep understandings of statistical hypothesis testing in introductory classes. However, in addition to being scarce in number, research reports on understanding of statistical hypothesis testing are limited in scope. Those studies that are done on a larger scale (with a large number of participants) are either focused on student error (Evangelista & Hemenway, 2002; Link 2002) or are focused on one aspect of understanding of statistical hypothesis testing: representation of sampling distribution (Hong and O'Neil, 1992) and *p*-value (Krauss & Wassner, 2002). Some studies have gone beyond student error in looking closely at student understanding of one aspect of statistical hypothesis testing on a small scale: understanding the relationship of sample size, effect size, and treatment size (Wilkerson & Olson, 1997) and the role of sampling distribution (Lipson, Kokonis, & Francis, 2003). On a large scale, Lane-Getaz (2007) studied student understanding of *p*-value and statistical significance.

Only a few studies have been conducted in which understanding of the entire concept of statistical hypothesis testing was examined but these were done with only a small number of participants (Aquilonius, 2005; Liu, 2005) some of whom were high school mathematics teachers, not introductory students (Liu, 2005). There have not been any studies that focus on student ability to recognize the role of context in studies that use statistical hypothesis testing and, as the conceptual analysis reveals, understanding of the role of context is important to the development of a deep, connected understanding of statistical hypothesis testing.

Based on this analysis, a large-scale study of introductory students' understandings of the "big picture" of statistical hypothesis testing would make a contribution to the development of a more complete picture of student understanding of statistical hypothesis testing. Furthermore, although there are some efforts to reform instruction in introductory courses, statistical instruction on the whole remains very similar across universities where introductory courses are taught largely through lectures with a great deal of emphasis on procedures (Garfield, Hogg, Schau, & Whittinghill, 2002; Shaughnessy, 1992). Of the students who take introductory statistics courses, a large number of them take these courses at the undergraduate level, at large universities. Therefore, a study of the understandings of students who have completed a traditional, university-level, introductory statistics course will inform future design of instruction on statistical hypothesis testing.

Given an incomplete picture of introductory students' understanding of the "big picture" of statistical hypothesis testing, a significant contribution to inform the design of instruction could be made by a large-scale study that addresses the following research question: *What are the understandings of statistical hypothesis testing held by students who have completed an introductory course in statistics at a large university?* Because information about student ability to implement the hypothesis testing process has been collected on a large scale, this study of introductory student understandings focuses more on examination of other aspects of understanding and the ways in which they do/do not connect. In other words, from the perspective of the *Adding It Up* model (Kilpatrick, et al., 2001), there is large-scale data that speaks largely to the *procedural fluency* strand and less strongly to the other strands or their connections. This argues that a study of student understanding should focus more strongly on the four other strands (*conceptual understanding*, *strategic competence*, *adaptive reasoning*, and *productive disposition*) as well as the connections among and within all five strands as they pertain to statistical hypothesis testing. And, in particular, with regard to the *productive disposition* strand, the study should focus on the degree to which introductory statistics students understand the value of statistical hypothesis testing, and believe it to be a useful method for inquiry. Therefore, this study addresses the following research sub-questions:

1. *What is the relationship between introductory students' understanding of the procedures and the concepts, logic, and uses of statistical hypothesis testing?*
2. *What are the understandings that introductory students have of the overall logic and reasoning of statistical hypothesis testing?*
3. *What are introductory students' understandings of the relationship between the method of statistical hypothesis testing and the context in which it is employed?*

This study addresses the research question and sub-questions outlined above through analysis of the understandings of statistical hypothesis testing held by university-level students who had completed (or nearly completed) a semester in an Introductory Statistics course at a large university. With a focus on procedures taught through lecture-style instruction (Garfield, Hogg, Schau, & Whittinghill, 2002; Shaughnessy, 1992), most university introductory courses use traditionally worded, well defined problems to assess student understanding. Although course assessments may give information about the *conceptual understanding, strategic competence, adaptive reasoning, and productive disposition* of students on some level, in essence these assessments provide information about the *procedural fluency* of the students. These assessments, therefore, can not be used to fully address the research questions and sub-questions. They only tell part of the story.

In order to collect information that addresses the research question and sub-questions on a large scale, (and that gives information about strands other than procedural fluency), a multiple-choice survey instrument was created and distributed to study participants in all sections of an introductory class (approximately 100 students). The questions on this multiple-choice survey instrument were designed to tap into those understandings of statistical hypothesis testing that are not typically assessed. The questions address the concepts, reasoning, logic, applicability, and role of context important to the overall, "big picture" of statistical hypothesis testing. The results of this assessment were analyzed and compared with those from a course exam, which focused entirely on statistical hypothesis testing. Together, the two quantitative assessments provided information about student understanding on a large scale.

In order to gain more insight into student understanding of the "big picture" of hypothesis testing, the multiple-choice survey and scores on third course exam were used to identify students who represent a range of performance patterns. Follow-up interviews were conducted with these students to provide valuable insight into their thinking. The mixed methods approach used in this study was valuable in filling an identified gap in the literature on student understanding of statistical hypothesis testing. In the chapters that follow, the details of the study are presented, including the process of instrument development; the results of both phases of the study (quantitative and qualitative); analysis of the data; the conclusions; and discussions of the implications, contributions, and limitations of the study.

CHAPTER 2

LITERATURE REVIEW

The conceptual analysis presented in the previous chapter highlights the complexity of statistical hypothesis testing as well as its utility in answering questions of interest. A review of research on understanding of statistical hypothesis testing indicates that students struggle in solving even well defined statistical hypothesis testing problems. In addition, a review of related research indicates that individuals often do not have well developed understandings of the various ideas and concepts involved. However, there are gaps in the literature that suggest a large scale study of introductory statistics students' understanding of this complex, useful concept is necessary to advance the field. Using large scale, quantitative data as well as small scale qualitative data, the study described in this dissertation explores the nature of students' understandings of this complex, useful concept. Existing literature on understanding of statistical hypothesis testing and its components is extremely important to this study as it informed both the development of the quantitative assessment instrument and the interpretation of the data collected in the quantitative and qualitative phases. Taken together, research on student understanding of statistical hypothesis testing, research on related concepts, and the results from this study contribute to the development of a description of student understanding of statistical hypothesis testing that will inform future design of instruction.

In this chapter, a review of literature will be presented. The chapter is organized into three sections: (1) a review of major findings from research on understandings of components of statistical hypothesis testing; (2) a review of findings from research on understandings of the “big picture” of statistical hypothesis testing; and (3) a summary of the findings and their impact on this study.

Research That Addresses Components of Statistical Hypothesis Testing

With the recent inclusion of statistical concepts into the school mathematics curriculum, and with increased emphasis on a statistically literate citizenry, educators have begun to explore issues related to the teaching and learning of statistical concepts and ideas. Thus, the field of “statistics education” is relatively young and the body of literature associated with this field is beginning to grow. The field of psychology, however, has made contributions to this line of study. Over the course of the past century, psychologists have been interested in the ways that humans reason and make predictions under conditions of uncertainty. Research in this area has implications for human understanding of statistical concepts and ideas as well as for the ways in which humans engage in statistical reasoning. Taken together, the studies conducted by psychologists and educators have contributed to a developing field of knowledge about student understanding of statistics and probability. Many of the findings are particularly relevant to statistical hypothesis testing and point to potential cognitive obstacles that must be overcome in order to develop a deep understanding of the method. Some of these areas include an understanding of the logic of proof as it applies to hypothesis testing, the ways in which samples should be chosen in order to make appropriate inferences, and the way that probability can be used to determine whether or not a sample is unusual.

Logic of Proof and Testing of Hypotheses

Human understanding of reasoning and proof as it applies to general hypothesis testing has been of interest to researchers for some time. Psychologists interested in studying the ways that individuals make decisions have studied human reasoning within the context of hypothesis testing through a series of tasks presented in psychology laboratories. Science educators interested in the development of the habits of mind associated with scientific inquiry have studied student reasoning in the context of testing hypotheses through experimentation in the classroom. Research in these two areas has revealed that individuals do not always rely upon the premises of logical proof that is the basis for general hypothesis testing.

Verification and Falsification

As was outlined in Chapter 1, a major premise of logical proof is that, unless exhaustive, verification of a hypothesis through generation of examples does not *prove* the hypothesis. On the other hand, generation of one contradictory case is all that is necessary to *disprove* a hypothesis. Studies have indicated individuals do not often apply these premises of logical proof to "test" a hypothesis. In order to test a hypothesis, humans often try to verify the hypothesis rather than attempt to falsify the hypothesis through generation of a counterexample.

One of the first researchers to investigate this phenomenon was Peter Wason (1967). In his seminal work with human reasoning Wason (as cited in Wason, 1967) presented his subjects with what has become known as the Wason 2-4-6 problem. In this problem, the researcher tells the subject that he/she is thinking of a rule that applies to groups of three whole numbers. The experimenter gives the subject one example of a triple that conforms to the rule. The first example presented to the subject is the triple 2, 4, 6. Given this triple, the subject must identify the rule. To determine whether his/her rule is correct, the subject produces a triple that conforms to his/her hypothesized rule. The experimenter tells the subject whether or not that triple conforms to the rule. The subject writes down his/her guesses and the reason for giving them. When the subject feels he/she has tried enough triples to identify the rule, he/she announces it. If the rule is correct, the subject has solved the problem. If the rule is incorrect, the experimenter tells the subject that he/she is not correct and the subject continues to generate triples and again tries to guess the rule (Wason, 1967).

This task proved difficult for Wason's (1967) subjects. The rule that dictates the formulation of the triples is that the numbers in each triple increase in magnitude. Few of Wason's (1967) subjects were able to correctly determine the rule. Critics of the task thought the rule was too difficult to determine. They claimed that subjects do not consider *numbers of increasing magnitude* to be a valid rule. However, in posing this problem to his subjects, Wason (1967) was not necessarily interested in whether the subjects were successful in determining the rule. Rather, he was interested the *reasoning* his subjects used.

In particular, Wason (1967) was interested in whether the subjects only tested triples that *confirmed* their hypothesized rule or whether they tested triples that could *falsify* the rule they had hypothesized. Only 21% of Wason's (1967) subjects correctly guessed the rule on the first announcement by varying testing triples that would confirm *and* falsify their developing hypotheses. Of the subjects who were unsuccessful, many only suggested triples that would *verify* their hypotheses. Even when their announced rule was deemed incorrect by the experimenter, the subjects continued to test triples that verified their original hypotheses. Wason termed this approach "verification bias" and found this to be a common approach used to do this task (Gorman, 1995).

It should be noted that the actions taken by Wason's subjects were typical of those taken by researchers engaging in scientific inquiry. Prior to formal testing of hypotheses, researchers generate hypotheses through a dynamic process of *informal* testing. The researcher generates a hypothesis, informally checks to see if it holds true for the data, and modifies the hypothesis based on information derived from the informal test. This process continues until the researcher feels confident that his/her hypothesis holds true. At this point, the researcher engages in more formal, scientific methods of testing his/her hypothesis. It could be argued that the subjects in Wason's study were engaging in this process of hypothesis generation. However, as noted by Wason, most of his subjects did not attempt to falsify their hypotheses at any point in the process.

Another task that elicits the verification bias is Wason's selection task. In this task, subjects are presented with four cards that show a vowel, a consonant, an even number, and an odd number (such as O, T, 6, and 9). The subjects are asked to determine which card(s) <u>must</u> be turned over to test the following claim: if a card has a vowel on one side then it has an even number on the other. To test the rule, the card showing the odd number and the card showing the vowel must be flipped. If the card with the vowel shows an even number on the other side, the rule has been verified. But, if it shows an odd number, then the claim has been proven false. Even if the card showing the vowel verifies the claim, the card showing the odd number must be flipped. If this card shows a vowel on the other side, the claim does not hold. If it does not show a vowel, then the claim is still valid. Thus, it is important to test cards that have the potential to *falsify* the claim (Evans & Newstead, 1995).

In Wason's study, many of the subjects tested the cards showing the vowel and the even number. Some simply tested the card that showed the vowel. Wason concluded that these subjects were operating under a "verification bias" as these cards would merely verify the claim. These subjects did not attempt to falsify the claim (Evans & Newstead, 1995).

Wason's research is seminal in that it sparked (and continues to spark) a great deal of study into human reasoning. Like Wason (1967), many researchers believe that this research indicates human tendency toward confirmation rather than falsification. Several researchers have attempted to find ways to reduce this tendency. Other researchers have offered other interpretations of these tasks and have, thus, modified the tasks in order to better understand the complexity of human reasoning. Reviews of research related to Wason's selection task have been written by Evans, Newstead, and Byrne (1993) and summaries of research on the 2-4-6 problem have been provided by Gorman (1995) and by Tweney and Chitwood (1995).

Wason's "verification bias" has implications for understanding how conclusions are drawn in scientific inquiry. In his research of student reasoning within the context of scientific hypothesis testing, Bady (1979) explored student understanding of the logic of proof. Specifically, Bady (1979) was interested in whether or not students understand that one confirming case does not prove a hypothesis but that one counter example can disprove a hypothesis. Furthermore, Bady (1979) was interested in the strategies students used to test a hypothesis. He wanted to know whether students sought to verify their hypothesis or if they attempted to falsify their conjectures (Bady, 1979).

In order to explore students' thinking, Bady (1979) asked 120 high school students (representing several grades within two different schools) to use various sets of data to evaluate a hypothesis. The students pretended to be a biologist who must test different hypotheses. The students were provided with different kinds of information to test their hypotheses. Having analyzed the responses provided by the students, Bady (1979) found that about half of the students failed to realize that a supporting case does not prove a hypothesis, a little over half realized that one counterexample disproved a hypothesis, and less than half attempted to use a falsification strategy to test a hypothesis.

Bady's (1979) results corroborate those reported for Wason's (1967) tasks and lend support to the notion that human beings are drawn toward verification as a method of proof. When asked to test a hypothesis, human beings do not attempt to use falsification. These studies indicate that human beings do not instinctively employ the laws of logic when testing hypotheses. Other studies, however, have indicated that the situation may be more complex.

Personal Belief and Experimentation

Further research on human reasoning in the context of hypothesis testing and experimentation has looked more broadly for explanations of the non-normative reasoning that human beings employ when making evidence-based decisions. This research has explored the factors that play a role in an individual's decision to retain or reject a hypothesis. While the notion of verification bias is present in human reasoning, research also indicates that factors associated with personal beliefs play a large role in the decision making process. Studies in this area show that personal theories and beliefs influence both the interpretation of data as well as the design of an experiment (Klaczynski, 2000; Kuhn & Dean, 2004).

One example that demonstrates the role of personal beliefs in reasoning is that of Klaczynski (2000). In his study of adolescent reasoning, Klaczynski (2000) highlights the fact that evaluation of evidence is highly influenced by the degree to which that evidence supports or refutes previously held, personal beliefs. According to Klaczynski (2000), when human beings are presented with data or information, they will employ one of two processes (or a combination) to evaluate the validity of this information. The individual may resort to an *analytic approach* in which he/she employs the laws of logic to evaluate the data as evidence that supports or refutes a given theory, belief, or hypothesis. On the other hand, the individual may employ a *heuristic* in which he/she judges the information based on the degree to which it is "consistent with stereotype-based and theory-based beliefs" (Klaczynski, 2000, p. 1348). Although both systems may be called upon to evaluate data, heuristics are easily activated as they require less cognitive demand (Klaczynski, 2000).

The degree to which data is congruent with personal theory affects whether or not an individual employs a heuristic process or an analytic process. When the data or information conforms to personal belief, individuals tend to employ a heuristic processes in evaluation of the data. In using this heuristic, the individual accepts the evidence as proof for his/her own belief. This is particularly true in adolescents for whom supporting evidence serves to not only preserve the personal belief but also to make conviction in that belief stronger (Klaczynski, 2000). Such reasoning is a form of verification bias.

It is when an individual is faced with evidence that is incongruent with personal beliefs that he/she is more likely to call upon analytic processes. Faced with information that provides evidence against personal theory, the individual must take a moment to think about that information and decide how to reconcile the contradiction. In this case, the individual will call upon analytic processes. However, if an individual turns to analytic processes to evaluate information, it does not guarantee that the individual will make the logical decision. He/she may still find a way to hold on to his/her personal theories (Klaczynski, 2000).

In his study of adolescent students, Klaczynski (2000) investigated the degree to which the students relied upon heuristic and/or analytic reasoning. He was particularly interested in how students reasoned when the information provided in the data related to personal beliefs. Klaczynski (2000) worked with a total of 130 students, some from grades seven and eight (early adolescent) and some from grades ten and eleven (middle adolescent). First, he gave the students a belief survey to determine their personal beliefs on social class and religion. Then, he presented the students with a series of scenarios in which a researcher had collected data to test a claim. The scenarios involved claims about social class and religion. The students were asked to indicate the strength of the researchers' conclusions and to provide justifications for their thinking. The conclusions offered by the researchers in the scenarios were supportive of a particular social class or religion. In addition, each conclusion offered by the researchers in the scenarios could be dismissed on any number of threats to validity introduced by the data collection and/or analysis. The personal theories survey was completed several more times throughout the questioning process (Klaczynski, 2000).

The results of Klaczynski's (2000) study supported the notion that adolescents use theory motivated reasoning to evaluate information that relates to personal beliefs. That is, they judge the validity of information based on the degree to which that information conforms to personal theories. Both early and middle adolescents displayed bias and heuristic reasoning to evaluate information that was congruent with their personal beliefs. In addition, both groups often used selective analytic reasoning to dismiss information incongruent with their personal theories, deeming that information to be "implausible". The beliefs survey indicated that when students were presented with information that did, indeed, support their personal beliefs, the students held stronger conviction in their theories. Deviations from normative reasoning and increased conviction in personal beliefs were more prevalent for the component of this study that addressed religion rather than social class. Finally, the results of the study indicated that though the older students possessed more advanced scientific/analytic reasoning skills than the younger students, they did not necessarily employ them more often than the younger students. Klaczynski (2000) reasoned that as adolescents get older they become more convinced of their personal convictions.

Klaczynski's (2000) study of adolescent reasoning indicates that human beings use verification bias and other forms of non-normative reasoning to evaluate information so that their conclusions are consistent with their personal beliefs. In their review of research on *scientific reasoning*, Kuhn and Dean (2004) found evidence that, in addition to impacting the ways that information is evaluated, personal beliefs impact the way that a researcher decides to collect data in an investigative study. Studies in scientific reasoning provide participants with data and ask them to make inferences about potential causal relations. Participants are monitored as they explore the data, search for patterns, and formulate inferences of causality. The researchers search for patterns of response among participants as they engage in the process of investigative inquiry. Collectively, the studies in this line of research have provided evidence that students' personal theories influence (1) the choices they make in deciding the features of the data to be used in analysis and (2) the inference rules or strategies they use in reaching a conclusion. In order to support their personal beliefs, students, therefore, (1) only select data for analysis that will confirm their personal theories and (2) hold identical forms of evidence to different standards of evaluation so as to provide evidence for their own theories (Kuhn & Dean, 2004).

The studies presented in this section indicate that human beings are not consistent in the way that they evaluate information to determine whether it supports or refutes suggested hypotheses. Research suggests that human reasoning in these situations is influenced by their own personal theories and beliefs. This phenomenon is found to hold, to some degree, for human beings at various stages of intellectual development, which suggests that personal theory influences reasoning more than developing understandings of logical proof does. Klaczynski's (2000) description of the analytic and heuristic approaches to reasoning indicates that unless humans make a conscious effort to reason according to the laws of logic, they will be subject to personal bias. According to Moshman (2004), it is metacognitive processes that dictate the degree to which humans employ logic when reasoning about a given situation. He suggests that, over time, individuals develop more sophisticated understandings of logic and, as they do so, they learn to monitor their reasoning so that they use this logic to evaluate their own and other people's reasoning (Moshman, 2004).

Though the studies presented in this section provide fairly convincing evidence that individuals use non-normative ways of reasoning with regard to hypothesis testing, it should be noted that none of the studies mentioned in this section investigated the instruction, if any, that individuals received on reasoning about logic and formal proof. Thus, these studies are limited. It may be that the misunderstanding and misconceptions identified in these studies can be linked to the way in which they learned about these concepts. Given a different set of learning experiences, these individuals might have reasoned differently about the tasks they were given.

Nevertheless, the results of these studies do, however, point to potential difficulties that individuals might have in developing strong understandings of statistical hypothesis testing. The notion that people test hypotheses through verification rather than falsification and that personal bias impacts the way that they both interpret and design research has implications for their understanding of and ability to use statistical hypothesis testing. The logic of indirect reasoning (falsification) is fundamental to the process and if this form of reasoning is not understood, then the individual may struggle to develop a complete understanding of statistical hypothesis testing. Understanding of and ability to use statistical hypothesis testing could be further hindered by personal beliefs. Personal theories may influence the way that an individual collects data, establishes a decision rule, and/or draws a conclusion or inference from that data.

Obtaining a Sample

As indicated by the Conceptual Analysis presented in Chapter 1 and the Framework for Assessing Understanding of Statistical Hypothesis Testing presented in Chapter 3, the method of data collection is an important consideration when using sample information to make inferences about the population in statistical hypothesis testing. The kinds of inferences that one may make from sample information are dictated by the way that the sample is collected. In order to apply inferential methods, one must be sure that the sample is not biased, that it was collected via a random process, and that the sample size is appropriate for drawing inferences. When these conditions are met, it is possible to draw inferences from a sample to the appropriate population. Research on student understanding of sampling has indicated that an awareness of the relationship between sample selection and the inferential process develops over time. Even when students have developed an awareness of important factors in sampling, it is difficult for them to coordinate these ideas in order to draw valid inferences about a population based on information from a sample.

In a review of research on student understanding of concepts from data analysis, Konold and Higgins (2003) outline some of the ways that middle grades students think about sampling. Often these students do not believe that inferences about a population can be made from analysis of a sample. Middle grades students believe that everyone in the population must be sampled. In addition, middle grades students do not recognize the potential for bias in sample collection (Konold & Higgins, 2003).

Konold and Higgins (2003) cite a study by Schwarz, Goldman, Vye, and Barron (1998) as an example to illustrate middle school students' lack of understanding of the relationship between sampling and inferential reasoning. The participants in the study were presented with a scenario about a school fair. In order to plan for the fair, students needed to estimate the number of children who would come to a booth at the fair. The participants were also presented with a scenario in which students needed to determine the number of boys and the number of girls at a school. Both scenarios required the participants to suggest methods to collect samples of 50 students from populations of 400 students. The students were encouraged to generate as many selection methods as possible (Schwarz, et al., 1998).

Overall, the participants suggested biased methods of sampling. In the booth scenario participants chose methods that employed a process of self selection, methods that sampled only those students who would be likely to come, and methods that would likely only include the students' friends in the sample. In the gender scenario participants suggested picking a sample of 25 girls and 25 boys. Some participants did attempt to employ random selection and suggested methods such as choosing 50 students without looking or choosing the first 50 students that come to school. However, these methods were still biased in some sense (Schwarz, et. al. 1998).

Konold and Higgins (2003), also cite a study by Jacobs (1999) in which late elementary students exhibited difficulty with the concept of sampling. In her study of student conceptions of sampling, Jacobs (1999) presented her fourth and fifth grade students with various situations in which samples of individuals were chosen to complete a survey that addressed a particular research question. The students were presented with sampling designs and the results that were obtained in using those designs. Some of the sampling methods introduced obvious biases and others did not. Jacobs (1999) asked her students to comment on the various strategies.

Jacobs (1999) observed as her students evaluated potential sampling schemes on the basis of fairness, practicality, believability of the associated results, or some combination thereof. In their concern for issues of fairness, many students chose methods that would result in self-selection. For these students, it was important that everyone be given the opportunity to complete the survey, whether they chose to complete it or not. Some students judged sampling methods based on whether or not the method was "efficient, easy to implement, confusing, or even plausible" (Jacobs, 1999, p. 245). Many students wanted to include everyone in the sample because those students "drastically underestimated the difficulties of asking everyone in large surveys" (Jacobs, 1999, p. 245). Some students used the results associated with a given sampling scheme to evaluate the method. If the results were indecisive and/or contradictory to the students' expectations, the methods were deemed ineffective. In addition, when presented with results from several studies that were contradictory, many students resorted to personal experience to evaluate the quality of samples used in each study, rather than considering issues of bias and randomness (Jacobs, 1999).

Together, the studies by Schwartz, et al. (1998) and Jacobs (1999) indicate that students do not understand the impact that sample bias and lack of random selection have on the quality of the results of a study. In their study of student understanding of sampling, Watson and Moritz (2000) found evidence that the ability to coordinate issues of sample bias, randomness, and sample size develops over time. Watson and Moritz (2000) analyzed third, sixth, and ninth grade students' responses to sampling problems.

In their analysis, Watson and Moritz (2000) identified six categories of developing concepts of sampling and hypothesized that these categories defined a developmental sequence for student understanding of sampling. Descriptors for the categories indicate that sensitivity to sample size and potential for bias increases with age. Younger students are satisfied with small samples (sample size fewer than 15) whereas older students require larger samples in order to draw an inference. Younger students are more likely to choose sampling methods that are biased and are not chosen via random selection whereas older students are more aware of these issues (Watson & Moritz, 2000).

As was the case for the studies presented in the previous section, care should be taken in assuming that the results here apply to all individuals. It may be that, given a different set of learning experiences, concepts of sampling will develop differently. However, the studies presented in this section indicate that an awareness of the connection between methods of sample collection and the quality of the inference that can be made from samples are important ideas upon which instruction should focus. Students' informal conceptions of sampling do not include an awareness of potential for sample bias, nor do they include an understanding of the role of sample size in attempting to make an inference based on a sample. Some students believe that it is not possible to get any information about a population using only sample information. Given the appropriate learning experiences, students can better coordinate the notions of sample bias, randomness, and sample size in evaluation of sampling methodology, but these issues remain a source of tension.

These findings have implications for the ways that introductory statistics students may or may not understand statistical hypothesis tests, and inferential methods in general.

If a student is not aware of the importance of randomly choosing a non-biased, representative sample large enough to make a valid inference, or if a student does not believe that one can make inferences from a sample, then that student probably will not have a well developed understanding of statistical hypothesis testing.

Using Probability to Determine if the Sample is Unusual

Once a sample is collected, it is analyzed to determine whether the sample is, indeed, unusual under the assumed null condition. This analysis lies at the very heart of statistical hypothesis testing. In order to determine whether or not the sample is unusual conditioned on the null hypothesis, one must take into consideration the variability that exists among randomly chosen samples. And, as was emphasized in Chapter 1, it is at this point that probability theory is useful. Under the assumed null condition, samples with a certain characteristic are considered to be *unusual* if the probability of obtaining such samples is small. Sampling distributions represent the distribution of all possible sample statistics for samples of a given size under the assumption that the null hypothesis is true. Thus, sampling distributions are useful for determining the probability that a randomly collected sample from the population described by the null hypothesis will be at least as extreme as the sample data. Therefore, in order to understand and appreciate the role of probability in the logic of hypothesis testing, one must coordinate understandings of sample variability, sampling distributions, and probability theory.

Research on human understanding of probability has been largely influenced by the work of Daniel Kahneman and Amos Tversky (e.g. 1971, 1972, 1982), who found support for a number of common heuristics that people employ when asked to determine the probability of a given event. In relying upon these seemingly intuitive heuristics, people are often led to make predictions and judgment decisions that are inconsistent with reasoning that is supported by probability theory. Kahneman and Tversky's (1971, 1972, 1982) findings sparked an entire body of research on human understanding of probability investigating ways that humans use (or fail to use) normative probabilistic reasoning when faced with situations that involve chance and uncertainty. Not only has this research indicated that people do, indeed, use non-normative patterns of reasoning, but it has also demonstrated that people are not consistent in their application of these non-normative patterns of reasoning across contexts. Additionally, it is difficult to replace these patterns of thinking with more appropriate methods of reasoning (e.g. Garfield & Ahlgren, 1988; Konold, 1995; Shaughnessy, 1992).

Research on human understanding of variability and, in particular, sampling variability, indicates that individuals do not have well developed understandings of the degree to which samples vary. These incomplete understandings may contribute to the non-normative patterns of reasoning identified in the probabilistic reasoning literature (e.g., Fong, Krantz, & Nisbett, 1986; Nisbett, Krantz, Jepson, & Kunda, 1983; Pollatsek, Konold, Well, & Lima, 1984; Reading & Shaughnessy, 2004; Well, Pollatsek, & Boyce, 1990). Not surprisingly, then, research has indicated that students struggle in the development of understandings of sampling distributions and their use in inferential reasoning (delMas, Garfield, & Chance, 2004; Saldanha & Thompson, 2002).

Such incomplete, undeveloped understandings of probability and sampling variability may impact the way that introductory statistics students understand statistical hypothesis testing. Therefore, it is important to examine this research.

Probability and Non-normative Ways of Reasoning

In their work with human reasoning and decision making under conditions of uncertainty, Kahneman and Tversky found that people “rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations” (Tversky & Kahneman, 1982, p.3). These heuristics are often useful but will, at times, lead to “severe and systematic errors” (Tversky & Kahneman, 1982, p. 3).

One heuristic identified by Kahneman and Tversky (1972) that seems particularly relevant to thinking about student understanding of statistical hypothesis testing is that of the *representativeness* heuristic. “A person who follows this heuristic evaluates the probability of an uncertain event by the degree to which it (i) is similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated” (Kahneman & Tversky, 1972, p. 431). In relying upon this heuristic, people will judge samples that resemble the population from which they are drawn to be more probable than those that do not. In addition, people will judge outcomes that “appear” to be randomly generated to be more likely than those that do not (Kahneman & Tversky, 1972).

In their study of the ways that tenth, eleventh, and twelfth grade students use probabilistic reasoning in sampling contexts, Kahneman and Tversky (1972) found that many of the students in the study relied on non-normative methods of reasoning. Each of 1500 students were given a questionnaire that contained questions about probability and sampling. They were asked questions such as the following: “All families of six children in a city were surveyed. In 72 families the exact order of births of boys and girls was GBGBBG. What is your estimate of the number of families surveyed in which the exact order of births was BGBBBB?" (Kahneman & Tversky, 1972, p. 432). (Note that, here, "B" represents a boy and "G" represents a girl.) Another item asked students to judge the likelihood that eight tosses of a fair coin would result in the following sequence: HTHTHTHT (where "H" denotes an outcome of heads and "T" denotes an outcome of tails). Kahneman and Tversky noted that these items elicited the representativeness heuristic in their participants' reasoning.

In the case of the birth order problem, both sequences are equally likely. The correct answer to the question is 72, or some number very close to 72. However, Kahneman and Tversky (1972) found that many of the students judged the first sequence to be more likely than the second. In the first sequence, GBGBBG, half of the children are boys and half are girls. In the second sequence, BGBBBB, five of the children are boys and only one child is a girl. With an equal number of girls and boys, participants in the study considered the first sequence to be more representative of the population. Therefore, the participants judged the first sequence to be more likely to happen than the second sequence. These students were employing the representativeness heuristic in their reasoning and it caused them to give an incorrect answer (Kahneman & Tversky, 1972).

In the case of the coin toss problem, Kaheman and Tversky (1972) found that, though the sequence HTHTHTHT contained an equal number of heads and tails, the participants thought this outcome was unlikely to occur. These participants did not think that the sequence appeared "random enough". Therefore, they did not think this was a likely result as it did not seem representative of the *process* by which the outcome was generated. Again, this form of reasoning is an example of reliance on the representativeness heuristic (Kahneman & Tversky, 1972).

Based on student performance on these and other questions, Kahneman and Tversky (1972) claimed that for many people, “a representative sample is one in which the essential characteristics of the parent population are represented not only globally in the entire sample, but also locally in each of its parts. A sample that is locally representative, however, deviates systematically from chance expectations: it contains too many alternations and too few clusters” (p. 435). As indicated by the examples presented above, this belief is problematic.

According to probability theory and the Law of Large Numbers, the size of a sample will affect the degree to which a randomly chosen sample will resemble the population from which it was drawn. The sample statistics associated with larger samples are more likely to approximate the population parameter than are sample statistics associated with smaller samples. When people rely upon the representativeness heuristic, however, they expect *any* sample, regardless of size, to be representative of the population. This non-normative reasoning leads to belief in what Tversky and Kahneman (1971) term the “Law of Small Numbers”. Operating in accordance with the Law of Small Numbers, the individual will make probability estimations that are larger than probability theory would predict, at least for small samples (Tversky & Khaneman, 1971).

Kahneman and Tversky (1971) caution that a belief in the Law of Small numbers and the notion of representativeness could lead to a common misconception associated with probability tasks: the gambler’s fallacy. In the coin toss example, if a sequence of outcomes does not contain an equal number of heads and tails then the next toss should produce a result that will “even out” the number of heads and tails. For example, if a four tosses of a fair coin results in HHHH, then by this line of reasoning, the probability of tossing a tail on the next toss is greater than that for a tossing a head. This belief in chance as a self correcting process is referred to as the gambler's fallacy (Kahneman & Tversky, 1972).

Another "fallacy" related to the representativeness heuristic is the *base-rate* fallacy. When reasoning according to this heuristic, "people ignore the relative sizes of population subgroups when judging the likelihood of contingent events involving the subgroups" (Garfield, 1995, p. 27). A typical question that elicits the base-rate heuristic involves judgments based on personality, rather than base-rate probabilities. For example, suppose that you meet a 45-year old male who is conservative and ambitious with no interest in politics. Which is more likely: (a) the man is an engineer or (b) the man is a lawyer? Based solely on the personality characteristics of this man, many individuals will claim that it is more likely he is an engineer. This same decision is made even when the individual is told that the man was randomly chosen from a population in which 30% of the people are engineers and 70% are lawyers. It has been argued that in ignoring base-rates, individuals use a form of representativeness heuristic in their reasoning. The man described is more representative of peoples' vision of engineers than of lawyers. In such cases, reasoning is based on an individual's experience with and/or knowledge of a given situation rather than on known proportions or probabilities (Shaughnessy, 1992).

A well studied problem used by Tversky and Kahneman (1982) that elicits the base-rate heuristic is the Taxi Problem. The problem is the following: "A cab was involved in a hit and run accident at night. There are two cab companies that operate in the city: a Blue Cab company, and a Green Cab company. It is known that 85% of the cabs in the city are Green and 15% are Blue. A witness at the scene identified the cab involved in the accident as a Blue Cab. This witness was tested under similar visibility conditions, and made correct color identifications in 80% of the trial instances. What is the probability that the cab involved in the accident was a Blue Cab rather than a Green one?" (Shaughnessy, 1992, p. 471). The correct answer is a probability close to 0.15. However, people tend to answer the question by offering a probability at or near 0.8. One explanation for this answer is that people ignore the base-rate information provided. Although only 15% of the cabs are Blue, people tend to place more confidence in the ability of the witness to make correct color identifications. They employ the base-rate heuristic and assume that the probability should be representative of the degree to which the witness is able to identify the color of vehicles (Shaughnessy, 1992).

Though the work of Kahneman and Tversky (1971, 1972, 1982) is well regarded, it has raised some questions concerning the degree to which the representativeness heuristic truly accounts for the responses provided by the participants in their studies. In particular, Konold (1989) hypothesized that the subjects in their studies were actually operating under different model of probability. "According to this model, referred to as the *outcome approach*, the goal in dealing with uncertainty is to predict the outcome of a single next trial" (Konold, 1989, p. 61). In other words, when individuals are asked to predict the probability of a given outcome, the number they state is indicative of the degree to which they believe that outcome will occur. A probability of 0.5 indicates the outcome could occur, a probability greater than 0.5 indicates that the outcome will occur, and a probability less than 0.5 indicates that the outcome will not occur. Individuals reasoning according to the outcome approach do not think stochastically about chance and probability. Instead they reason deterministically (Konold, 1989).

Konold (1989) designed a study to test his hypothesis and found support for his conjecture. When asked to estimate the probability of an event, Konold (1989) found that often individuals assume that the question is asking for a prediction about the outcome for the very next trial. These individuals are interpreting and answering the question using an outcome approach. Konold (1989) also found that when individuals employ an outcome approach to predict events, they do so to the degree to that causal factors can be identified in the situation. For example, when Konold (1989) asked participants what a forecaster means when she claims a 70% chance of rain, some participants told him that, since 70% is greater than 50%, the forecaster is saying it will rain. When the question was rephrased to ask what could be concluded from a situation where it didn't rain on a day that the forecaster predicted a 70% chance of rain, some participants said that the forecaster must have been wrong. Other participants offered explanations that relied on causal factors, such as a change in the wind. These responses indicate that the participants were not thinking of 0.7 as a probability generated from repeated trials. Rather, because the number given was greater than 50%, the number indicated that it would, indeed rain the next day (Konold, 1989).

Konold (1989) also tested his hypothesis on responses to the Taxi problem and, in so doing, found that often the base-rate fallacy can be attributed to the outcome approach. When participants were asked the probability that the cab was blue, some of them asked whether they were supposed to give a number. These participants wanted to simply answer with a color – the color of the cab they believed to be involved in the accident.

This response is indicative of a view of probability as prediction of the next outcome. Another participant claimed that it didn't matter how many cabs in the city were blue and that the information about the proportions was extraneous to the problem. The participant reasoned that the cab in the story is the only cab that should be considered and, since the witness is mostly correct (greater than 50-50), the cab must have been blue. Again, this response is indicative of the outcome approach. The participant interpreted the 80% reliability rate to mean that the witness is correct. No other information was required because there was only one situation under consideration, the one in which the witness is involved (Konold, 1989).

In a later study Konold, Pollatsek, Well, Lohmeier, and Lipson (1995) demonstrated that, often, the gambler's fallacy can also be attributed to the outcome approach. In their study participants were asked which is most likely to occur on 5 flips of a fair coin and were given the options: HHHTT, THHTH, THTTT, HTHTH, or all are equally likely. The participants were then asked which was least likely and were given related options: HHHTT, THHHTH, THTTT, HTHTH, or all are equally unlikely. The participants were further probed to explain their answers. The researchers found that some participants were using an outcome approach to answer the question. For example, one participant answered both questions correctly, but when she was asked to state the *probability* that the HTHTH sequence would occur, she stated that because it is *possible* for the coin to produce that sequence, the probability is 0.5. For her, a probability of 0.5 indicates that the outcome could occur. This answer indicates that the participant interprets probability as a statement of what will happen on the next trial. Again, this is an example of reliance on an outcome approach to probability (Konold, Pollatsek, Well, Lohmeier, & Lipson, 1995).

The different approaches to probability may be due to the kinds of learning experiences these individuals with probability concepts. Given different learning experiences, individuals might develop different understandings of probability that may or may not be correct. However, whether it is the case that individuals reason from an outcome approach or according to a representativeness or base-rate heuristic, it seems to be well documented that people generally use non-normative ways of reasoning in situations that involve probability. In all cases, the reasoning employed indicates a lack of consideration of the variability that exists in sampling situations. In relying on the representativeness or base-rate heuristic an individual does not acknowledge the fact that samples may vary and assumes all samples should resemble the population from which they are drawn. In using the outcome approach to probability, an individual does not interpret probabilities as indicative of the uncertainty variation that is involved in repeated events and/or sampling.

The concept of sampling variability is central to statistical hypothesis testing and to inferential statistics in general. Thus, it is important that individuals have an understanding of the ways in which samples vary. Research in this area, though, indicates that humans do not have well developed understandings of the concepts and ideas involved.

Sampling Variability

The development of statistical concepts and methods was motivated by a desire to make decisions based on data. Given that there is a great deal of variability associated with data, statistical methods and concepts were developed to deal with this variability. Hence, the concept of variability is central to the practice of statistics. Unfortunately, research has indicated that people are uncomfortable with variability. People do not have strong understandings of the variability that exists within a sample nor do they understand the variability that exists among samples generated by a process of repeated sampling. This difficulty in understanding variability has an impact on the degree to which individuals predict the probability that a given sample would be collected from a specified population.

Research on student understanding of variability within a sample has indicated that students do not have strong understandings of variance as a measure used to describe the distribution of values. Given the distribution of values in a data set (or sample) students will refer to means and proportions to describe that data set. They do not refer to concepts associated with variability, such as the spread of the data to describe the distribution. Reading and Shaughnessy (2004) claim that these findings are not at all surprising considering statistics education places a great deal of emphasis on measures of center but not on measures of variation. In addition, students are rarely asked to identify sources variation and/or find ways of visualizing or measuring variation in a data set. Studies have indicated, however, that students' reasoning about variation in a data set is influenced by the way those students think about centering and clustering. As students focus on clusters as well as centers, they can begin to develop stronger understandings of variance within a distribution (Reading & Shaughnessy, 2004).

The discomfort students have with variance within a sample extends to a discomfort in dealing with the variability that occurs between samples. As students reason about samples, they experience a tension in dealing with sample variability and sample representativeness (Reading & Shaughnessy, 2004). These ideas seem somewhat contradictory. How can different samples drawn from the same population vary from each other while remaining representative of that population? Complicating matters is the issue of sample size. How does sample size impact the degree to which sample characteristics vary from and/or are representative of the population? The ability to reconcile these tensions is important for developing an understanding of sampling variability. Research on human reasoning about these issues has uncovered some interesting ways that humans reason in conditions where sampling variability, sample size, and representativeness should be considered.

In a study of understanding of sample variability, Nisbett, Krantz, Jepson, and Kunda (1983) demonstrated that they degree to which individuals are aware that there is *potential* for sample characteristics to vary impacts subsequent evaluations of the likelihood of a given sample. In this study, participants were told to pretend they had landed on an island that had not been well explored. On this island, the participants were told that they came in contact with several creatures. Some of the creatures were found by themselves, some in groups of three, and some of groups of twenty. The participants were told some of the characteristics of these creatures. For example, the participants encountered birds that were blue, natives that were obese, and a new element that burned green when heated. Given varying sizes of samples of these objects, the participants were asked the degree to which they believed other members of the groups associated with those objects had the same characteristic (color, obesity, burn color) (Nisbett, et al., 1983).

Nisbett, et al. (1983) found that that degree to which participants believed a particular characteristic (color, obesity, burn color) to be variable had an impact on the size of the sample the participants thought was required to generalize that characteristic to the population. If a characteristic (such as obesity) was considered to be variable in a population, then a larger sample would be necessary in order to make any inferences to the population about that characteristic. Nisbett, et al. (1983) concluded that individuals only employ the representativeness heuristic in particular domains – those for which the characteristic under question is deemed variable. Therefore, under certain conditions, people *do* rely upon the Law of Large Numbers and are hesitant to draw inferences from small samples. That is, individuals rely upon the Law of Large Numbers (not the law of small numbers) when they understand the sampling process as well as the variability associated with the sample space (Nisbett, et al., 1983).

In addition, there has been some evidence that training in the Law of Large Numbers can improve performance on tasks that typically evoke reasoning according to the law of small numbers. Fong, Krantz, and Nisbett (1986) studied the effect using pre- and post-tests. All of the participants in the study were given a pre-test that contained problems that typically elicit reasoning using the law of small numbers. The participants were separated into four groups. The control group did not receive training on the Law of Large Numbers. The other three groups received some form of training. One group was presented with a description of the Law of Large Numbers. Another group was presented with scenarios accompanied by explanations of how the Law of Large Numbers can be used in reasoning about those scenarios. The final group was presented with both forms of training. After the training had occurred, a post-test was administered to measure the impact of this training on performance. Fong, Krantz, and Nisbett (1986) found a positive correlation between the amount of training and the degree to which performance increased. As the amount of training increased, the ability to correctly apply statistical reasoning also increased (Fong, Krantz, & Nisbett, 1986).

In another study of understandings of sampling variation Well, Pollatsek, and Boyce (1990) found that though people understand that sample means of larger samples are more likely to resemble the population mean (the Law of Large Numbers), they do not understand the impact of sample size on *variability* of the sample means. In a series of studies, the researchers provided the participants with different scenarios in which participants were told the population means. The participants were then asked questions about likelihood of obtaining various sample statistics for a variety of sample sizes (Well, Pollatsek, & Boyce, 1990).

For example, the participants were told the following: “When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 9 inches”(Well, Pollatsek, & Boyce, 1990, p. 293). Different groups of participants were then asked different questions. One group was asked whether the sample mean of a sample of 25 men would be closer to 5 feet, 9 inches than would the sample mean of a sample of 100 men (accuracy version). Another group was told that, for a year, samples of 25 men were taken every day at one location and samples of 100 men were taken at another location. The participants were then asked to state whether they thought there was a greater number of days for which the sample means were 6 feet or more at the first location or at the second location (tail version). A third group was asked the same question but instead of asking for the relative likelihood that the sample means were “6 feet or more” the participants were asked to make predictions that the sample means would be “between 5 and 6 feet” (center version). Throughout the study, participants were asked similar questions situated in other contexts (Well, Pollatsek, & Boyce, 1990).

Well, Pollatsek, and Boyce (1990) found that, for each of the contexts presented, the participants did well on the accuracy version, reasonably well on the center version, and did not do well on the tail version. The researchers concluded that individuals have difficulty recognizing that sample means vary and that sample size plays a role in the degree to which samples vary. This finding has implications for the way in which individuals understand sampling distributions (which will be discussed later in this chapter). Well, Pollatsek, and Boyce (1990) conducted a follow-up study in which they demonstrated to participants the way that sampling distributions are created. Unfortunately, even after the participants had observed this demonstration, they continued to have difficulty reasoning about the way that sample size affects variability of the mean. Well, Pollatsek, and Boyce (1990) concluded that people often believe that “extreme scores are more likely to occur in large samples (which is true) and that, therefore, the averages of large samples will be more variable (which is not true)” (p. 310).

The results of the study conducted by Well, Pollatsek, and Boyce (1990) suggest that the concept of sample variability is difficult to understand. In fact, one could argue that the representativeness heuristic, in general, is indicative of a struggle to coordinate the notion that samples may vary. The concepts of sample representativeness and sample variability are difficult to connect when making estimates of the likelihood of a given sample. Bruce (1991) illustrates this notion well with the following statement:

> One of the keys to mastering statistical inference is balancing these two ideas, interpreting more precisely the meaning of "likely" in each. Because they are contradictory when seen in a deterministic framework, students may over-respond to one or the other depending on the context. Over-reliance on sample representativeness is likely to lead to the notion that the sample tells us everything about a population; over-reliance on sample variability implies that a sample tells us nothing. Finding the appropriate point on the continuum between the two extremes is complex and needs to take into account confidence level, population variance, and sample size. For a given confidence level and population variance, the effect of sample size relates closely to the representativeness/variability continuum: the larger the sample, the more likely it is to be representative of the population. Smaller samples are more likely to vary.
>
> (Bruce, 1991, section 1.1, ¶ 5)

Managing the tension between sample representativeness and sample variability is a real concern in inferential statistics and practicing statisticians are forced to deal with this issue on a regular basis (Reading & Shaughnessy, 2004).

Several researchers have begun to study the development of understanding of variability and have subsequently proposed frameworks that outline various levels and kinds of understandings of variation (e.g. Torok & Watson, 2000; Watson, Kelly, Callingham, & Shaughnessy, 2003). Building on the work done by previous researchers on the construction of developmental hierarchies for understanding, Reading and Shaughnessy (2004) conducted a study whose goal was the refinement of these hierarchies. Reading and Shaughnessy (2004) showed students two different bags filled with 100 lollipops each. Each bag contained red, blue, and yellow lollipops. The students were told that the first bag contained 50 red, 30 blue, and 20 yellow pops and that the second bag contained 70 red, 10 blue, and 20 yellow lollipops. Given the proportion of red, blue, and yellow lollipops in each bag, the students were asked to determine how many red pops they would expect to get if they were to take a sample of 10 lollipops from each bag. They were then told that six different students chose a sample of 10 lollipops from each bag. They were asked how many red pops they would expect to appear in each of the six samples taken from each bag. The participants were then asked how many red pops they would expect if those six students had chosen samples of 50 lollipops each. Finally, they were asked how many red pops they would expect if 40 students each chose a sample of 10 lollipops from each bag (Reading & Shaughnessy, 2004).

In their analysis of student responses, Reading and Shaughnessy (2004) found evidence that understanding of the variation in sample statistics develops with respect to two different hierarchies. One hierarchy addresses the ways that students are able to *describe* variation. Within this hierarchy, responses at the first level indicate that the individual focuses *either* on middle values (population proportion) or on extreme values in predicting the number of red lollipops in a given sample. For example a student might claim that numbers close to the population proportion are likely *or* that numbers further from the population proportions are not likely. Second level responses indicate that the student is paying attention to both middle *and* extreme values in his/her reasoning about the likelihood of sample proportion. At the third level, student responses indicate an awareness of “anchoring” from some value, not explicitly stated as the center. For example, a student’s response might indicate that he/she understands that the values should lie “on either side” or that the values should fall a given distance from an extreme value. At the fourth (and final) level, student responses clearly indicate that the student is “anchoring” from the center (population proportion). Responses in this category explicitly refer to the center and what is happening around the center. Such responses also consider the effect of sample size on the spread of potential results (Reading & Shaughnessy, 2004).

The second hierarchy identified by Reading and Shaughnessy (2004) addresses the explanation students give for the *cause* of the variation that occurs. At the first level, students attribute variation to extraneous causes such as whether the lollipops were well mixed. At the second level, causality is attributed to the frequency of reds in the bag. For example, a student might justify his/her answer based on the fact that "there are a lot of reds in there". At the third level, student responses indicate that the *proportion* of reds (not just the number) influences the number of reds in a given sample. Finally, fourth level responses discuss the likelihood of getting a given sample based on the proportion of colors in the bag (Reading & Shaughnessy, 2004).

Given the focus that is typically focuses on measures of center with little attention to variability, the hierarchies identified by Reading and Shaughnessy (2004) indicate that development of understandings of sample variation can occur students (1) become aware of the potential for sample characteristics to vary and (2) recognize that potential for a spread of sample statistics is determined by the size of the sample and the population parameter. Both hierarchies point to the notion that students need to begin to understand that repeated sampling produces a distribution of sample statistics. Within that distribution, some values are more likely than others. The statistical concept that connects these ideas and that attaches probability values to sample statistics is the sampling distribution. Given reports of student difficulty in coordinating the idea that samples can be representative and will still vary in ways that are dependent upon the sample size, it stands to reason that students would struggle to develop strong understandings of sampling distributions. Research has indicated that this is the case.

Sampling Distributions

Sampling distributions represent the distribution of sample statistics and are used to attach probabilities to samples for specified populations. Research has indicated that it is difficult for students to understand that sample statistics vary and that this variability can be measured and represented in a distribution. In order to help students better understand sampling distributions, some educators have suggested that instruction should provide students with the opportunity to *construct* sampling distributions. Because the process of drawing samples from a population is laborious, researchers have developed software that allows students to engage in the construction of sampling distributions. Using this software, students can construct sampling distributions for a variety of sample sizes, population parameters, and number of samples chosen. In constructing these sampling distributions, it is hoped that students will understand that sample statistics vary and that the size of the samples impacts the overall shape of the distribution. As is the case with use of any technology, the question remains as to whether the students using the technology are making appropriate connections between the technology and the way that it demonstrates important concepts in statistics. More importantly, the question remains as to whether the technology improves students' ability to engage in the statistical reasoning necessary for statistical inference (delMas, Garfield, & Chance, 2004).

In reviewing research, including their own previous work in this area, delMas, Garfield, and Chance (2004) found that students do not develop the desired understandings of sampling distributions, even when sampling distribution simulation software is incorporated into instruction. In their review, delMas, Garfield, and Chance (2004) found that, though students could recite a series of facts related to sampling distributions, use of technology did not help them to develop reasoning skills that would enable them to use sampling distributions in statistical inference. Rubin, Bruce, and Tenney (cited in delMas, Garfield, & Chance, 2004) found that, even after use of technology to construct sampling distributions, students were not able to coordinate an understanding of the relationship between sample size, sample representativeness, and sampling distribution. Students had trouble understanding that as sample size increases, samples are more likely to resemble the population but as sample size increases, the sampling distribution looks *less* like the distribution of values in the population. A study by Hodgson (cited in delMas, Garfield, & Chance, 2004) found that the use of simulations could actually contribute to the development of misconceptions.

Having done this review of research, delMas, Garfield, and Chance (2004) attempted to make improvements upon the way in which technology enhanced simulations are used to develop strong and useful understandings of sampling distributions. Having identified a series of common misconceptions, the researchers endeavored to design instruction that would address these misconceptions and replace them with more desirable understandings. Some common misconceptions associated with sampling distributions of sample statistics include the following:

- The sampling distribution should look like the population (for n>1)
- Sampling distributions for small and large samples have the same variability
- Sampling distributions for large samples have more variability
- Don't understand that a sampling distribution is a distribution of sample statistics
- Confuse one sample (real data) with all possible samples (in distribution) or potential samples
- Confuse implications of The Law of Large Numbers with the Central Limit Theorem
- The mean of a positive skewed distribution will be greater than the mean of the sampling distribution for samples taken from this population

(delMas, Garfield, & Chance, 2004, p. 8)

The researchers used the identified misconceptions to inform design of an instructional activity addressing sampling distributions. delMas, Garfield, and Chance (2004) tested the students before and after the activity to determine whether or not the misconceptions had been replaced with more desirable understandings. Based on the results of the pre- and post-test comparison, the researchers modified the activity by breaking it into smaller pieces. After having observed students during that broken activity and upon analysis of the results of a pre- and post-test comparison, the researchers again modified the activity. They included a reflection piece where students were led in a discussion that forced them to think more deeply about the activity. The results of this improvement were assessed using the same pre- and post- test comparison (delMas, Garfield, & Chance, 2004).

Much to their surprise, the researchers did not find an improvement of performance over the three versions of the activity. In fact, the same misconceptions appeared in many of the students. Many students did not understand that while large samples resemble the population, the distribution of sample means for large samples does *not* resemble the population distribution. The students believed that as the sample size increased the distribution of sample means would look more like the population. Many students also confused variability in a sampling distribution with variability within the population and/or sample. delMas, Garfield, and Chance (2004) concluded that students need more experience in working with densities in the distribution of a sample and more experience working with densities in the distribution in the parent population. delMas, Garfield, and Chance (2004) also reported that students tend to rely on informal understandings when reasoning about sampling distributions. Students do not take a moment to stop and think about the situation. Given that sampling distributions are very complex concepts, and given that students do latch onto informal understandings, delMas, Garfield, and Chance (2004) recommended that students have more exposure to the various concepts and relationships involved.

Noting the complexity of the concept of sampling distributions and the difficulty that students face in connecting sampling distributions to statistical inference, Saldanha and Thompson (2002) designed a study to explore the *development* of student understanding about sampling distributions. Given that students struggle to manage the tension between sample representativeness and sampling variability and given that some people approach probability with an outcome approach in which the repeatability and distribution of sampling is not recognized, Saldanha and Thompson (2002) designed a teaching experiment in which they focused on the development of sampling ideas. Instruction was, therefore, designed to “support [students’] conceiving sampling as a scheme of interrelated ideas including repeated random selection, variability among sample statistics, and distribution” (Saldanha & Thompson, 2002, p. 259). Instructional design emphasized that (1) random selection can be repeated under like conditions and (2) patterns emerge from collection of samples and these allow for estimates of the likelihood of particular sample statistics on the basis of relative frequencies (Saldanha & Thompson, 2002).

Twenty-seven eleventh and twelfth grade students participated in the teaching experiment. Instruction began by engaging the participants in an examination of studies that used sampling in their analysis. The instructors then asked the students to determine what fraction of the time they might expect results like those reported in the studies. The students then used a computer simulation to inform their initial responses to these questions. Given the parameters identified in the studies, the students used the computer to create sampling distributions so that they could "see" the fraction of the time the results reported in the studies actually occur. It was hoped that through engagement of the activity, students would recognize the importance of sample size in determining the shape of the sampling distribution (Saldanha & Thompson, 2002).

In talking with the students, Saldanha and Thompson (2002) found two different conceptions of sampling that affected the degree to which the students were able to develop deep understandings of sampling distributions. Students who had "developed a multi-tiered scheme of conceptual operations centered around the images of repeatedly sampling from a population, recording a statistic, and tracking the accumulation of statistics as they distribute themselves along a range of possibilities" (Saldanha & Thompson, 2002, p. 261) were better able to answer questions posed during instruction. These students had developed what Saldanha and Thompson (2002) referred to as a *multiplicative conception of sample*. This conception allowed the student to simultaneously think about the sample as a part of the whole "in terms of the whole" (Saldanha & Thompson, 2002, p. 266). Students who had developed this conception were able to think about relative frequencies and likelihoods associated with various values for sample statistics (Saldanha & Thompson, 2002).

On the other hand, students who were not able to think of the process as a multi-tiered scheme of operations struggled to keep track of the number of people in the sample, the number of samples chosen, and the distribution of sample statistics as a distribution. These students had what Saldanha and Thompson (2002) referred to as an *additive conception of sampling*. These students "view a sample simply as a subset of a population and view multiple samples as multiple subsets" (p. 265). These students are not able to think about a distribution of sample statistics and their relative frequencies within that distribution (Saldanha & Thompson, 2002).

The development of a multiplicative conception of sampling is of particular importance to hypothesis testing. The decision to reject or fail to reject the null hypothesis is based on the degree to which a sample is deemed unusual under the null condition. In order to determine whether or not the sample is unusual, one must have the ability to think of sampling as a random, repeatable process for which sample statistics may be calculated and represented in a distribution. The collection of sample statistics creates a sampling distribution which attaches probabilities to values of the sample statistic. These probabilities give information as to whether or not the sample is unusual. If students have an additive rather than a multiplicative conception of sampling, they may not be able to fully connect and apply the essential ideas involved in statistical hypothesis testing. These students may continue to struggle with the tension between sampling representativeness, sampling variability, and sample size and this may impact the development of their understanding of statistical hypothesis testing.

In another study of the development of student understanding of sampling distributions, Lipson, Kokonis, and Francis (2003) identified four stages of development through which students must progress in order to understand the relationship between computer simulations and statistical inference. The researchers presented a small group of students with a claim that the post office delivers 96% of letters on time. The students are told that a journalist decides to test the claim and sends out 59 letters. The journalist reports that 52 (88.1%) of the 59 letters were delivered on time. He claims that the post office is incorrect. The students were subsequently asked to use the statistical software to better understand the journalist's claim. The software created a sampling distribution of the proportion of letters delivered on time from 200 samples of 59 letters in a population where the proportion delivered on time was 96%. As the students worked with the software, the researchers guided them in their activity. Lipson, Kokonis, and Francis (2003) collected observational data on the students' progression through the activity and upon analysis, identified four stages important to the development of understanding of the software and its relation to the *particular* problem at hand.

In the first stage, the *recognition stage*, students began to understand the representation shown on the computer screen and its relation to the situation. Students first began to understand that the distribution on the screen was a distribution of sample statistics. They then started to understand that sample statistics vary and that the variance is represented in that distribution. They then began to understand how to use the sampling distribution to determine how often a given sample statistic occurs. In the second stage, the *integration stage*, students began to use the sampling distribution to decide whether the sample containing 52 letters delivered on time is a likely result. In the third stage, the *contradiction stage*, students began to recognize the contradiction created in the *integration stage*. That is, at this stage, students explicitly acknowledged that the sample the journalist collected was unlikely if the post office was, indeed, correct. Lipson, Kokonis, and Francis (2003) noted that this stage was difficult for students. Though they recognized the small likelihood of the sample, the students didn't seem to think this was a problem. The students tended to give practical rather than statistical reasons for this contradiction. The researchers had to help the students move to stage three so the students could move to the final stage, the *explanation stage*. In this stage, students were able to offer an explanation for the contradiction. In case of this particular problem, the students needed to recognize that the post office may not be correct in their claim to deliver 96% of the letters on time. However, the researchers note that this final stage is difficult for students to reach. In summary, though, the researchers claim that the computer software was valuable in helping students to progress through these stages (Lipson, Kokonis, & Francis, 2003). Though these stages were important for the students to progress through this *particular* instructional activity, they provide some guidance in thinking about instructional design for developing ideas associated with sampling distributions and statistical inference.

The research by Lipson, Kokonis, and Francis (2003) extends that by delMas, Garfield, and Chance (2004) and Saldanha and Thompson (2002). Collectively, the studies highlight the difficulty that students have with understanding the concept of sampling distributions. In addition, the studies indicate that students may struggle to understand the role of sampling distributions in determining whether a given sample is unusual or not for a specified population. In particular, students have difficulty coordinating sample size, sample variability, and sample representativeness in reasoning about the likelihood of a given sample in a known population. The results also indicate that these difficulties persist even after instruction has been modified to specifically target these difficulties.

Though these studies explored student understanding of the likelihood of a sample, they did not explicitly connect probability to sampling distributions. Research on student understanding and interpretation of *p*-value indicates that students do, indeed, struggle to understand the connection of probability to sampling distributions and the role that each plays in statistical hypothesis testing.

Understanding and Interpretation of p-values

In statistical hypothesis testing, the *p*-value gives the probability of obtaining the collected sample statistic (or more extreme) under the condition that the null hypothesis is true. Therefore, *p*-values are useful in determining whether or not the collected sample is unusual if the null hypothesis is the correct description of the population. Because they provide a measure of the likelihood of the result under the assumed null condition, *p*-values are often reported in journals. It is important, therefore, that introductory statistics students understand what the *p*-value represents and how to interpret the *p*-value within the context of the study. They should understand the relationship between *p*-value and sample size and should know how to coordinate this information in interpretation of studies that report a high (or low) *p*-value for their results. Research has suggested, however, that this concept is not easily understood by students and/or researchers.

Krauss and Wassner (2002) studied understanding of *p*-value on a large scale. They developed a questionnaire that outlined various (incorrect) interpretations of *p*-

value and asked participants to decide if the interpretations were true or false. They distributed their questionnaire to 113 students, instructors, and scientists representing 6 German universities. The questionnaire asked the following question:

> Given that t = 2.7, df = 18, p = 0.01 in an independent means t-test comparing means of two populations, determine if the following are true or false (more than one or none of them may be true)
>
> A. You have absolutely disproved the null hypothesis (that is, there is not a difference between the population means)
> B. You have found the probability of the null hypothesis being true
> C. You have absolutely proved your experimental hypothesis (that there is a difference between the population means)
> D. You can deduce the probability of the experimental means being true
> E. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision
> F. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of the occasions
>
> (Krauss & Wassner, 2002, p. 2)

Though each of the statements is false, many of the respondents, including the instructors, claimed that at least one of the statements was true (Krauss & Wassner, 2002). The results indicate that students and instructors alike do not have strong understandings of *p*-value. They do not know what the *p*-value reported in a study represents, nor do they know how to interpret a low *p*-value within the context of a given study.

Other studies have explored students' and researchers' understandings of the relationships between and among *p*-value, statistical significance, sample size, and treatment effect. Wilkerson and Olson (2001) designed a questionnaire that addressed these relationships and distributed it to 52 graduate students. Of the 52 participants, 20 were pursuing a PhD, 14 were pursuing an EdD, and 16 were pursuing a master's degree. Only one respondent realized that when two studies report the same (low) *p*-value, the study showing the greatest evidence for a treatment effect is the one that used a *smaller* sample. The sampling distributions of statistics for smaller samples are more spread and obtaining a sample statistic that lies in the extreme gives more evidence for a treatment effect. In addition, Wilkerson and Olson (2001) found that most of the respondents thought that sample size affected the probability of making a Type I error. This is not correct. The level of significance (α) is the probability, in the case where the null hypothesis is true, of committing a Type I error. It is a decision criterion set prior to data collection. Sample size does not impact the probability of committing a Type I error. Approximately half of the respondents, however, did recognize that sample size impacts the probability of committing a Type II error, in the case that the null hypothesis is not correct. However, the lack of understanding of the relationships between and among sample size, treatment effect, and statistical significance is problematic in that it may affect the ways in which individuals interpret the results of their own or other people's studies (Wilkerson & Olson, 2001).

The misunderstandings identified in Wilkerson and Olson's (2001) study do not simply reside in graduate students. In their study of the statistical understandings held by members of the American Educational Research Association (AERA), Mittag and Thompson (2000) found similar results with regard to understandings of *p*-value, sample size, statistical significance, and treatment effect. Mittag and Thompson (2000) developed a questionnaire containing a variety of claims (both correct and incorrect) about a collection of statistical concepts. For each of the claims, participants were asked to use a Likert scale to indicate the degree to which they agreed with each claim. Some of claims addressed participants' understandings of the relationships between and among *p*-value, sample size, statistical significance, and treatment effect. Like the graduate students in Wilkerson and Olson's (2001) study, most of the respondents did not agree that "statistically significant results are more noteworthy when sample sizes are small" (Mittag & Thompson, 2000, p. 18).

The reports of student difficulty in understanding *p*-values and statistical significance indicate that instruction on these ideas has not been successful. As instructors begin to modify their approach, it will be necessary to evaluate the degree to which the new approaches are successful. In recognition of a need for an assessment instrument that could be used across studies of instruction, Lane-Getaz (2007) endeavored to create such an instrument. Her goal was to create an instrument could be used (a) to assess student understanding of *p*-value and statistical significance; (b) to provide a consistent, reliable measure of student performance that can be used across studies of student understanding of *p*-value and statistical significance; and (c) to establish well defined learning goals for these concepts. Lane-Getaz (2007) created an instrument that addressed the understandings and misunderstandings of *p*-value and statistical significance that are currently highlighted in the literature. Several iterations of the assessment were piloted with feedback from students, educators, and statistical experts. With the final two iterations of her instrument, Lane-Getaz (2007) found strong content- and weak construct-related validity. Lane-Getaz (2007) cautions, however, that the test should not be used to assign grades in an introductory statistics course as the reliability of the test was low. The test can, however, be used across research studies to evaluate the success of instruction on *p*-value and statistical significance (Lane-Getaz, 2007).

In her study, Lane-Getaz (2007) also reported the results of student performance on her assessment instrument. She found that students often associated the *p*-value with the probability that the null hypothesis is false. In addition, students interpreted a study reporting a low *p*-value to indicate that the results were *caused* by chance rather than that they *could* be attributed to chance. Lane-Getaz (2007) also noted that, students do not understand that random sampling is necessary for valid inferences to be made from a sample to population.

Together, the studies of understanding of *p*-value and statistical significance found in the literature indicate that these are difficult concepts for students. Understanding of these concepts requires coordination of understandings of the relationships between and among sample size, sample variability, treatment effect, and probability. These concepts are essential to statistical hypothesis testing may impact the degree to which introductory statistics students understand "the big picture" of statistical hypothesis testing.

Summary

The studies presented in this section provide evidence that individuals do not have strong understandings of some of the components of statistical hypothesis testing. Lack of understanding of these components could impact the degree to which introductory statistics students understand the "big picture" of statistical hypothesis testing.

For example, if introductory statistics students are prone to use verification or reliance on personal beliefs rather than falsification to test hypotheses, they may struggle to understand the use of indirect reasoning in statistical hypothesis testing. In fact, personal conviction might impact the degree to which introductory statistics students value the use of statistical hypothesis testing to test hypotheses. If a student has a strong enough belief in the feasibility of the alternative hypothesis, for example, he/she might consider the hypothesis to have been proven correct if the sample chosen confirms that hypothesis. In this case, the student would not feel there was a need to engage in formal testing, such as statistical hypothesis testing, because the sample proved the hypothesis. It may be difficult for them to understand that test statistics cannot *prove* a hypothesis.

With regard to sampling, the research presented here indicates that students rely on informal ways of thinking about methods of data collection. When engaged in a task that requires sampling to make an inference about the population, students often suggest biased and non-random methods of data collection. Their suggestions indicate difficulty in understanding sample representativeness and sample variability. Understandings of these concepts are essential to an understanding of the ways that statistical hypothesis testing can be used to draw inferences about a population using information obtained from a sample. In particular, these understandings are essential to an understanding of the role of probability in determining whether or not a sample is unusual under the assumed null condition.

With regard to probability, research has indicated that humans rely on heuristics that lead to non-normative ways of reasoning about the likelihood of specified outcomes. These non-normative ways of reasoning highlight the challenge that individuals face in coordinating the notions that samples should be representative of the population, that samples can vary, and that sample size has an impact on the degree to which samples are representative and/or variable. The struggle that students face in reconciling those ideas extends to their understandings of sampling distributions. Students often struggle to understand the way that sampling distributions connect all of these ideas to probability. In addition, students struggle to understand and interpret *p*-value, statistical significance, level of significance, and treatment effect. These understandings are essential to an understanding of statistical hypothesis testing.

Given that research on individuals' understanding of various components associated with statistical hypothesis testing indicates that people do not have strong understandings of these concepts and ideas, it is reasonable to expect that introductory statistics students do not have strong understandings of the entire method and its uses. And, in fact, there is evidence to support the notion that students do not have well developed understandings of the overall process of statistical hypothesis testing. Though studies in this area are scarce, a review of that literature will be presented in the next section.

Research on Understanding of Statistical Hypothesis Testing

As indicated in Chapter 1, there are very few studies that address introductory statistics students' understandings of the overall concept of statistical hypothesis testing. The research that *has* been done suggests that introductory statistics students (and their teachers) struggle to develop deep, connected understandings of the concept. They often make errors in implementing the steps required to solve traditionally worded, well-defined hypothesis testing problems. In addition, when presented with ill-defined statistical hypothesis testing problems, individuals are not able to apply the method to solve those problems. Though limited, these studies indicate that introductory statistics students do not develop strong understandings of the “big picture” of statistical hypothesis testing.

Student Error and Persistence after Instructional Intervention

A review of literature on student on student understanding of statistical hypothesis testing reveals several articles providing anecdotal evidence that students make mistakes when performing the steps of a statistical hypothesis test. However, there is one systematic study of student errors in the literature. This study, conducted by Link (2002), examined student errors in solving statistical hypothesis testing problems for a large number of students.

After noticing that students struggle to solve problems that involve statistical hypothesis testing, the instructors at LSU-Shreveport designed instruction to focus on what they identified as “a six-part procedure” for doing statistical hypothesis testing problems. Students were instructed to do the following six steps in order to solve statistical hypothesis testing problems: (1) set up hypotheses, (2) find and state the critical value, (3) construct a probability statement that connects probability to the test statistic, (4) state the observed value of the test statistic, (5) compare the observed value of the test statistic and the critical value in order to decide whether to reject the null hypothesis, and (6) state the *p*-value (Link, 2002).

In using this six-part procedure as a guide, Link (2002) then performed an analysis of six exams given in one of 6 sections of an introductory statistics class. In total, 295 student exams were analyzed. Five of the exams used in the analysis were final exams and the sixth exam was a regular class exam. In the analysis of student work, Link (2002) took note of the part(s) of the process for which students made errors and noted the kind of error that was made. Link (2002) found several common errors illustrated by the students' work. Some of the students stated the population parameter incorrectly, some used the sample statistic in their statement of the hypotheses, and some chose the correct population parameter but did not state the hypotheses using the correct inequality. Many students made errors in their statement of the test statistic (critical value), most often in cases for which the test was two-tailed. Finally, most students struggled to construct the appropriate probability statement. On the other hand, most students were able to make the correct decision concerning whether to reject the null and were able to use a table and/or graphing calculator to give the *p*-value (Link, 2002).

The results of Link's (2002) study indicate that students make errors in virtually every step of statistical hypothesis testing. It should be noted that these mistakes persisted even after instruction was designed specifically to focus on the very parts of the method that were to be assessed. In giving the students a series of steps that were used in application of *any* statistical hypothesis test (e.g. tests of mean, proportion, etc.), it was hoped the students would be successful on the exam. Unfortunately, this form of instruction did have the desired effect (Link, 2002).

Other attempts to design instruction aimed at improving student performance on statistical hypothesis testing problems have had similar results. In their work with college students, Evangelista and Hemenway (2002) identified several difficulties that students have with statistical hypothesis testing. These include:

> - Inability to distinguish a test of hypothesis situation from other situations such as estimation or finding probabilities
> - Failure to recognize the population parameter to be tested and whether more than one population is involved
> - Difficulty specifying the null and alternative hypothesis and determining the rejection region
> - Confusing the sample and the population…
> - Difficulty interpreting their conclusion …
> - Poor understanding of the reasoning …even if they can procedurally do all the textbook exercises
>
> (Evangelista & Hemenway, 2002, p.2)

In an attempt to focus students' attention on the similarities among specific statistical hypothesis testing procedures (test for a proportion, test for a mean, one-tailed, two-tailed), Evangelista and Hemenway (2002) designed a jigsaw activity for their students. Expert groups worked with one "type" of statistical testing problem. Then, new groups composed of one member from each expert group were formed. The members of these new (jigsaw) groups compared and contrasted the method used to solve each problem. It was hoped that, as a result of this activity, students would recognize the logic and general steps associated with all statistical hypothesis tests. However, at the end of the activity, the researchers were not convinced that this instructional design had improved student performance on statistical hypothesis testing problems (Evangelista & Hemenway, 2002).

Another study that examined the effects of instructional design on student performance on statistical hypothesis testing problems was conducted by Hong and O'Neil, Jr.(1992). In this study, the researchers spent time talking with expert statisticians and found that experts frequently refer to graphs of sampling distributions and rejection regions in their explanations of how one draws or interprets a conclusion to a statistical hypothesis test. As a result of this discovery, the researchers were interested in whether or not presentation of diagrammatic models would improve student performance on statistical hypothesis testing. In addition, Hong and O'Neil (1993) were interested in determining whether an instructional approach that focused first on concepts and then procedures would be more effective than one that focused on concepts and procedures concurrently (Hong & O'Neil, 1993).

The researchers assigned a total of 56 students to one of 4 treatment groups. Two groups received instruction that taught concepts, then procedures. One of those groups was presented with diagrams during instruction and the other group was not. The other two groups received instruction that taught both concepts and procedures simultaneously. One of those groups was presented with diagrams during instruction and the other group was not. The diagrams used in the presentations illustrated the sampling distribution, the location of the critical value, and the location of the observed value of the test statistic. In addition, rejection regions were shaded. All instruction was delivered through lessons on a computer (Hong & O'Neil, 1993).

Hong and O'Neil (1993) found that those students who received instruction that used diagrams performed better on the post-test than those who did not. In addition, students who received conceptual and then procedural instruction performed better than those who did not (Hong & O'Neil, 1993). Though these results seem to demonstrate that instruction can improve performance, the study does not report the kind of questions that were used to assess understanding on the post-test. Therefore, the question still remains as to whether the students really developed a deep understanding of the concepts or if they merely improved performance on items that assessed procedural competency.

Understandings Other Than Procedural Fluency

Because reports of student error in performing statistical hypothesis tests have called into question the nature of the understandings held by introductory statistics students, a few studies have been conducted that examine these understandings a deeper level. These studies have revealed some interesting insights into the ways in which individuals understand statistical hypothesis testing.

In her dissertation study, Aquilonius (2005) explored the nature of introductory statistics students' understandings of statistical hypothesis testing beyond mastery of procedures. In order to study the ways in which students think about statistical hypothesis testing, Aquilonius (2005) observed as eight pairs of community college students solved problems that involved statistical hypothesis testing. During her first meeting with each pair of students, Aquilonius (2005) asked the students to talk out loud as they solved a series of traditionally worded, well-defined statistical hypothesis testing problems. At their second meeting, each pair of students was given problems which were accompanied by hypothetical student solutions. These hypothetical student solutions were designed by Aquilonius (2005) to be representative of common student errors she had observed in her instruction. The participants were asked to evaluate the students' work and make corrections where necessary. Each meeting was videotaped for subsequent analysis (Aquilonius, 2005).

Upon analysis of the videotapes and student work, Aquilonius (2005) found that the participants frequently confused population and sample means. As they "talked aloud" and critiqued student work, the students in the study did not readily refer to sampling distributions. Aquilonius (2005) concluded that the students did not use sampling distributions in their reasoning about the problems. In the “talk aloud” and critique of student work, Aquilonius (2005) also noted that the students did not have well-developed, conceptual understandings of *p*-values. Upon finding the *p*-value for given problem, the participants did not give a reason for their decision to reject or to fail to reject the null hypothesis. In making their decision, the students often recited a rule they had been taught. In addition, the participants did not seem to connect probability theory to the work they were doing. Aquilonius (2005) noted that the participants seemed to equate randomness with representativeness and did not fully understand the need for a hypothesis test. There were no discussions of the Central Limit Theorem nor were there appeals to probability as the students discussed their work with each other (Aquilonius, 2005).

The results of Aquilonius’ study indicate that though students may be able to perform the operations required in a statistical hypothesis test, they don’t necessarily have well developed conceptual understandings. As they “talked aloud” and discussed student work with each other, the pairs of students in her study typically gave procedural explanations for their work by reciting well-rehearsed rule statements. As Aquilonius (2005) noted, the students did not connect probability theory and/or the notion of sampling variability to their work. These ideas are central to statistical hypothesis testing and it seems reasonable to assume that if they are not understood students will think of statistical hypothesis testing as an algorithm rather than a tool to apply logical and probabilistic reasoning to decision making.

In her dissertation study of *teachers’* understandings of statistical hypothesis testing, Liu (2005) also found that individuals struggle in the development of a deep, connected understanding of statistical hypothesis testing. Liu's (2005) study was part of a larger project conducted by a team of researchers led by Patrick Thompson. The goal of the project was to explore the relationship between multiplicative and stochastic reasoning. The project was comprised of five studies and Liu (2005) reported on the last study in the project. The first four studies used teaching experiment methodology to investigate the development of high school students' understandings of sampling and statistical inference. The study by Saldanha and Thompson (2002) reported earlier in this chapter was one of the studies. Based on the analysis of data and insights collected in the first four studies, the research team designed a seminar for teachers to think about the statistics they teach and the way that they teach it (Liu, 2005).

Eight teachers, each with experience teaching high school statistics, participated in the seminar. The teachers also completed a pre- and post-interview. Liu (2005) analyzed data collected during the interviews and during each session of the seminar. In particular, Liu (2005) was interested in teachers' conceptions of probability and statistical inference in the context of confidence intervals and statistical hypothesis testing (Liu, 2005).

In her analysis of the data, Liu (2005) reported several interesting findings. Liu (2005) found that teachers struggled in the development of an understanding of the unusualness of samples. Teachers struggled to understand the way that sampling distributions are constructed and used to determine whether or not a sample is unusual given a specified population. Liu (2005) also found that teachers did not fully understand the logic of hypothesis testing. For example, they did not understand that application of hypothesis testing requires a commitment to the alternative hypothesis in hopes that the null will be rejected. In addition, when faced with a small *p*-value, teachers often used non-normative ways of reasoning about the implications of that small value. These non-normative ways of reasoning were usually based on personal belief about the situation prior to conducting the hypothesis test. For example, when told that a particular study reported a low *p*-value, one teacher questioned whether the sample was randomly chosen. Another refused to reject the null because there wasn't overwhelming evidence against it. Additionally, Liu (2005) found that the teachers did not understand statistical hypothesis testing as a tool and, thus, did not know when it should be used to answer a question of interest (Liu, 2005).

Summary

In summary, the studies presented in this section give support to the claim that individuals struggle to develop deep, connected understandings of statistical hypothesis testing. Introductory statistics students make mistakes in every step of the procedure when solving well-defined, traditionally worded statistical hypothesis testing problems. Even after having received instruction targeted at improving student performance, those errors persist. As studies by Liu (2005) and Aquilonius (2005) indicate, even if students do perform well on these problems, it does not necessarily mean that these students have a deep understanding of the method and its uses. It should be noted that the studies conducted by both Liu (2005) and Aquilonius (2005) were done with a small number of participants. And, in one study those participants were teachers with experience in teaching statistics. Additionally, Aquilonius' (2005) was focused on student explanations of well-defined, traditionally worded hypothesis testing problems. She did use non-traditional questions to gain insight into her students' understanding. Therefore, more exploration into introductory statistics students' understandings of statistical hypothesis testing is needed to advance the field.

Chapter Summary and Implications for Study

Overall, the studies in this chapter highlight a series of misconceptions and informal, non-normative ways of reasoning that individuals rely on when making predictions and evaluations of likelihood under conditions of uncertainty. In addition, studies indicate that individuals are not inclined to test hypotheses using indirect methods. Rather, individuals tend to test hypotheses using verification techniques and that that these methods are often influenced by personal beliefs and experiences. These findings have implications for the ways in which individuals are able (or not able) to develop strong understandings of the "big picture" of statistical hypothesis testing.

Results of the few studies conducted to examine student understanding of statistical hypothesis testing provide some support for the notion that a student's ability to perform the steps of traditionally worded, well-defined problems does not necessarily mean that he/she has well-developed understandings of the logic and concepts that support the use of statistical hypothesis testing in real world contexts. A few studies of individual understandings provide evidence to support this claim (e.g. Aquilonius, 2005; Liu, 2005). However, those studies were limited in scope and were conducted on a small number of individuals. In one case the subjects were teachers. In addition, those studies did not assess the degree to which individuals understand the role of context in performing a statistical hypothesis test. Therefore, more research into the understandings of introductory students is necessary. From the perspective of *Adding It Up* (Kilpatrick, et al., 2001), with respect to statistical hypothesis testing, more large scale research is needed on the degree to which introductory statistics students have *conceptual understanding*, *strategic competence, adaptive reasoning*, *and productive disposition* (in the sense that students understand the value of the method) as well as on the relationship of these proficiencies to *procedural fluency*. The study described in this dissertation addresses those needs.

Findings from research on student understanding both of statistical hypothesis testing and of various components of statistical hypothesis testing was used to inform the design of the large scale study described in this dissertation. The findings presented in this chapter informed both the creation of the multiple-choice instrument and the analysis of quantitative and qualitative data collected in this study. The methodology, results, and conclusions of this study are presented in the remaining chapters.

CHAPTER 3

RESEARCH METHODOLOGY AND DESIGN

Analysis of the theoretical and empirical literature on student understanding of statistics suggests the need for a large scale study of the understandings about hypothesis testing that students develop after experience in a standard introductory statistics course. The study outlined in this chapter employed both quantitative and qualitative methods to address that objective. The combination of research strategies allows for general claims about the understandings of a large number of students and more substantive claims about a small number of students.

This chapter describes the methodology for the study in three sections. The first section presents the rationale and key components of the chosen mixed methods research methodology. The second section outlines the research design used in this study. Included in that section is a description of the participants; the setting; the timeline; and the instrumentation, data collection, and data analysis associated with each phase of the study, data collection, and data analysis for the study. The third section presents a summary of the chapter.

Research Methodology

Researchers can choose from a broad range of methods for studying student understanding in knowledge domains that are both complex and highly structured. Each particular research method allows for collection of data of a certain kind. So each method has both limitations and benefits associated with it. In general, methods that obtain information from large numbers of students do not provide in-depth information about understandings of individual students. On the other hand, methods that explore individual student understanding in depth are not feasible for gathering information about large numbers of students. Therefore, every researcher must decide what data on student understanding would be most useful to his/her study and choose the appropriate method(s) accordingly. In the design of a study of understandings about statistical hypothesis testing, several options were considered in searching for a strategy that would provide both generality and depth of findings.

One common method for surveying student understanding in a domain is to use a collection of multiple-choice questions. Such instruments are easily distributed to large numbers of students, the responses are easy to score, and the results are easy to summarize. However, results of multiple-choice testing seldom provide a complete and convincing picture of student understanding. Both the choice of questions on the survey and the options provided to respondents limit the kind of information about understanding that is obtained. Thus, while a multiple-choice survey approach to study of student understanding can provide data for a large number of students, it is limited in the depth with which data may be analyzed.

On the other end of the research methods continuum, a more qualitative approach using clinical interviews can help the researcher develop a more complete picture of student understanding. "Clinical interviews can give more information on depth of conceptual understanding, because oral and graphical explanations can be collected and clarifications can be sought where appropriate" (Clement, 2000, p. 547). While this kind of research method will provide deeper insight into student thinking, data analysis is more labor intensive. Thus, studies that utilize clinical interviews are generally conducted on small numbers of students. For that reason, the generalizability of findings is always questionable.

Both multiple-choice surveys and individual clinical interviews (and methods that lie in between these ends of the quantitative/qualitative continuum) have benefits and limitations associated with them. As a result, many researchers are turning to mixed methods approaches in studies of social phenomena. In using both quantitative and qualitative methods to study a phenomenon, researchers are able to create "an analytic space that doesn't necessarily resolve the tensions but rather uses them – in respectful conversation – to probe more deeply …"(Greene, 2001, p. 252). For example, in a study of student understanding, one may use what Greene (2001) terms a *complementary* mixed-method design in which a quantitative study is followed by interviews. In this case, "results from one method are intended not necessarily to converge with but rather to elaborate, enhance, illustrate, or clarify results from the other" (Greene, 2001, p. 253).

In order to fill an identified gap in the field of statistics education—the need for a large scale study of overall understandings about hypothesis testing—and to capitalize on the power of clinical interviews to provide in-depth portraits of student understanding, a complementary mixed-method research design was used. Large scale information on student understanding was obtained using quantitative assessments in the quantitative phase and follow-up interviews were conducted in the qualitative phase to gain more insight into student thinking on a small scale.

Research Design

In using a mixed methods approach, this study of student understanding of statistical hypothesis testing employed both quantitative and qualitative phases of data collection and analysis. The methods used to collect this information included surveys and follow-up interviews. These methods were used to study student understanding of statistical hypothesis testing among students enrolled in introductory statistics at a large university. The surveys (or assessments) were used in the quantitative phase while the interviews were conducted during the qualitative phase.

Methods

Assessments given in the normal activity of university level introductory courses provide some information on student understanding on a large scale and were used in this study. However, these exams tend to focus their assessment on procedural fluency. To identify common patterns of student understandings about statistical hypothesis testing that are *not* traditionally assessed on course exams (such as *conceptual understanding* and *adaptive reasoning*), a multiple-choice assessment instrument was created and administered to a large sample of students in an introductory university course. The results from use of the multiple-choice instrument as well as student performance on the course exam covering statistical hypothesis testing were used to identify a strategically chosen sample of 11 respondents to engage in follow-up interviews that yielded deeper insight into student thinking than was provided by the multiple-choice instrument and scores on course exams alone. These three data sources provided information about the three research sub-questions as illustrated in Table 3.1.

Table 3.1

Research Sub-question and Data Sources

Research Sub-Question	Course Exam (Large scale)	Multiple-choice Assessment (Large scale)	Interview (Small scale)
1. What is the relationship between introductory students' understandings of the procedures and the concepts, logic, and uses of statistical hypothesis testing?	√	√	√
2. What are the understandings that introductory students have of the overall logic and reasoning of statistical hypothesis testing?		√	√
3. What are introductory students' understandings of the relationship between the method of statistical hypothesis testing and the context in which it is employed?		√	√

Taken together, the data collected in the two phases of the study provide a strong basis for descriptive claims about student understanding of statistical hypothesis testing and its applications.

Participants

The participants in this study were students enrolled in eight sections of a one-semester undergraduate introductory statistics course at a large Research I university. The typical student in this class has declared a major in one of the social sciences and was taking the course because it is required for that major. The course is not open to mathematics majors, and only rarely will a student from the "hard" sciences or engineering register for it.

The study participants were chosen because they are similar to groups of students who enroll in introductory statistics classes across the country. As was noted in Chapter 1, increasingly diverse groups of students with various goals and motivations are enrolling in introductory statistics courses (ASA, 2005). Students have different backgrounds and take the course for different reasons (e.g., to satisfy a major requirement or to learn methods they will use in discipline-based research). The participants in this study share the characteristics of those described in the GAISE report. For this population, the GAISE report has recommended course goals that focus on the development of conceptual understanding, statistical literacy, and an ability to engage in statistical thinking. These students should become *statistically educated* (ASA, 2005). It is reasonable, therefore, to examine whether they develop the reasoning and thinking skills described in that report. In particular, this study provides information as to whether these students finish the course with desired understandings of statistical hypothesis testing.

Setting

The study was conducted at a large, public research university located in an urban area of the eastern part of the country. Enrollment at the school is approximately 35,000 with approximately 25,000 undergraduate and 10,000 graduate students. The introductory statistics course is taught in sections of approximately 30 students which meet three times a week for 50-minute class sessions. The instructors of the course, offered by the department of Mathematics, included one lecturer and three graduate assistants working on degrees in mathematics, computer science, and statistics respectively. Mathematics and other "hard" science majors are required to take the calculus based version of this course and do not register for this course. In addition, many departments on the university campus have developed their own versions of an introductory statistics course, which their majors take in lieu of this course. However, students may take STAT 100 to prepare for a course in their major.

Instruction in the course is largely lecture based. The design of the lectures closely follows the presentation of concepts and material provided in the quite traditional introductory statistics text, *Statistics: Principles and Methods* by Johnson and Bhattacharyya (2006). Topics covered include descriptive statistics, probability, probability distributions, sampling distributions, point estimation, confidence intervals, and hypothesis testing. The course has a less ambitious syllabus than other introductory courses and somewhat more emphasis on probability than other courses. Grades are assigned based on student performance on exams, quizzes, homework, MINITAB projects, and a final exam. Students are allowed and encouraged to use calculators. Students are provided with necessary statistical tables and formula sheets, which they may use on assessments.

The course described here is typical of many offered across the country. As was mentioned in the first chapter, many of the introductory statistics courses are taught similarly via a lecture format with a great deal of emphasis on procedures (Garfield, Hogg, Schau, & Whittinghill, 2002; Shaughnessy, 1992). Although recent recommendations for different forms of instruction that promote deeper understanding (e.g. ASA, 2005; Cobb, 1992; Garfield, 1995; Moore, 1997;Wild & Pfannkuch, 1999) have encouraged instructors to change their teaching methods, the implementation of those recommendations has been limited. According to the GAISE report (ASA, 2005) and a study of introductory statistics teaching practices by Garfield, Hogg, Schau, and Whittinghill (2002), most introductory courses are changing only in the use of technology. Students are allowed to use calculators, and computer use is encouraged in student assignments. The course from which participants were drawn for this study has made these changes but it remains lecture based with a focus on procedures. It is, indeed, similar to many other introductory statistics courses across the country. Therefore, the analysis of student understanding in this study provides insight into the ways that many other introductory students understand statistical hypothesis testing.

Time Line

Prior to data collection, time was dedicated to development of a framework for assessing understanding, construction of the multiple-choice instrument, and for piloting of both phases of the study.

Both the framework for assessment and the multiple-choice items were developed during the 2006-2007 academic year. Working closely with the director of the dissertation, versions of each piece were created and sent to a statistician for feedback. Modifications were then made based on that feedback. This process continued throughout several iterations.

The multiple-choice items and follow-up interviews were then piloted with one section of the introductory statistic course in the spring of 2007. Fifteen multiple-choice items were piloted. Roughly half of the items (7 questions) were included on “Form A” and roughly half (8 questions) on “Form B”. Follow-up interviews were then conducted with 3 students who represented a range of performance patterns on the multiple choice assessment and overall class performance. In the follow-up interview, students were asked to explain their thinking on the 7 (or 8) items they answered as well as to explain their thinking on additional items presented to them for the first time during the interview.

The quantitative phase of the pilot study provided information on the frequency with which individual distractors were chosen. The qualitative phase provided insight into participant interpretation of the language used in each item and whether he/she was able to answer the item using a process of elimination that did not require deep understanding. This information was useful to further refine the stems of each item, the answer choices, and the language used. In addition, the pilot study provided valuable information about timing. In the pilot study, the participants spent approximately 10-12 minutes answering 7 (or 8) multiple-choice questions in class. In the follow-up interview, participants began to fatigue after having been asked to explain their thinking for longer than an hour. This information was useful in determining both the number of multiple-choice questions that could be used in an assessment scheduled to take 20-25 minutes of class time and the number of questions that could be asked in an hour long follow-up interview.

Modifications were made to the multiple-choice instrument and to the interview protocol during the summer and fall of 2007. The quantitative phase of the study was conducted in December, 2007 and the qualitative phase during December, 2007 and January, 2008.

Quantitative Phase: Data Collection and Analysis

In order to provide information on student understanding on a large scale, data was collected via a course exam and the newly created multiple-choice assessment.

Though participants were required to take the course exam (as part of the course, itself), they were not required to take the multiple-choice assessment for the course. Students from all 8 sections of STAT100 were asked to volunteer to participate in the study, which required them to take the multiple-choice assessment and to release their exam scores to be used in the study. Each multiple-choice assessment was marked with a number. A third party liaison was the only person who had a link between student names, multiple-choice assessment numbers, and student scores on the third course exam. The liaison prepared a list that linked the numbers on the multiple-choice assessment to exam scores for analysis. Participants' assessment numbers were included in a lottery to win one of ten, $20 gift cards. The liaison helped to distribute these gift cards.

The course exam was designed by the course coordinator with input from the course instructors. Though various forms of the exam were created, each form was similar in nature and focused entirely on statistical hypothesis testing. Questions on the exam were similar to those worked in class and/or for homework. The following problem was used on a previous exam and is similar to those asked on the current exam:

Figure 3.1

Sample Item, Course Exam

> A survey of 50 university juniors finds their average credit card debt to be \$3900 with standard deviation \$900 while a survey of 50 university seniors finds their average credit card debt to be \$3500 with standard deviation \$500. Perform a hypothesis test with $\alpha = .01$ to determine if there is a significant difference between the two mean credit card debts. Include all of the steps.

Notice that the problem provides students with the appropriate summary measures of the data. It is a well defined problem, similar to those traditionally found in introductory statistics textbooks. Students have access to calculators, statistical tables, and formula sheets. Although the problem does assess the various strands of proficiency identified in Chapter 1 to *some* degree, the focus is on student ability to apply the algorithm. Therefore, the course exam provided a measure of *procedural fluency*, at least with respect to well-defined, traditional statistical hypothesis testing problems.

The multiple-choice assessment consisted of 14 questions designed to assess understandings of the statistical hypothesis testing that are not traditionally assessed on course exams (such as *conceptual understanding* and *adaptive reasoning*). In order to construct such an instrument, a framework for assessing understanding was developed. This study, addressed the degree to which students have deep understandings of the conceptual and logical foundations as well as the uses of statistical hypothesis testing, given that they are traditionally assessed only on ability to apply the procedure. Therefore, the multiple-choice assessment provided a measure of the degree to which students understand the foundations and conceptual underpinnings of the algorithm. The assessment framework is organized according to the actions involved in conducting a study that uses statistical hypothesis testing. The assessment items were written to address the underlying concepts and theoretical principles that support those actions. The Framework for Assessing Understanding is presented in Figure 3.2.

Figure 3.2

Framework for Assessing Understanding of Statistical Hypothesis Testing

Category	Process	Understanding Assessed
Recognizing Applicability (RA)	1. Identify situations where statistical hypothesis testing is an appropriate strategy for addressing a research question. 1.1. The research question can be answered by analysis of some measure of the population. 1.2. The research question is formulated so that it addresses a well defined population for which it is only feasible to study a sample from that population. 1.3. There exist two conflicting, contradictory hypotheses that can answer the research question. 2. Identify the value of using statistical hypothesis testing to answer the research question of interest.	• Indirect reasoning will be employed and, therefore, two competing hypotheses are needed. • It is necessary to quantify the question so that probability may be used in determining whether or not the sample is unusual under the assumed null condition. • Statistical hypothesis testing provides a means of "answering" a research question about a population given information from a sample and uses probabilities to quantify the uncertainty necessarily associated such an inference
Generating Statistical Hypotheses (GH)	1. State null and alternative hypotheses so that: • the hypotheses are contradictory; • the alternative hypothesis is consistent with what the researcher would like to prove. 2. State hypotheses so that they address the measure identified in the research question. 3. State hypotheses so that a practical decision can be made.	• Indirect reasoning will be employed and, therefore, two competing hypotheses are needed. • Writing the hypotheses to indicate a one- or two-tailed test will address the practical needs of the researcher. If directionality is important for practical interest, hypotheses should be written accordingly (for a one-tailed test). If not, then hypotheses are written to test only for equality or inequality.
Decision Rule (DR)	1. Determine a decision rule for rejection of the null hypothesis that is based on • choice of a test statistic • probability • the degree to which the researcher would like to be confident in his/her conclusion • what is considered unusual for the null condition but expected in the alternative condition	• A "cut point" is necessary to determine whether to reject the null hypothesis or not. This decision takes into account the probability associated with the sample, given the null hypothesis. • A "cut point" determines the risk of Type I error the researcher is allowing for – if the null is true, and the researcher rejects the null, this is the probability he/she is wrong.

Collect a Sample (CS)	1. Recognize the need for a non-biased, random sample while understanding the notion that samples may vary. 2. Recognize the way that samples must be obtained so that one may, indeed, answer the question of interest based on the data collected.	• The way in which a sample is chosen will affect the nature of the inference that can be drawn, including the population to which that inference can be applied. • In statistical hypothesis testing, samples must be unbiased, random, and large enough. • Samples are expected to vary. Larger samples are more representative of the population.
Analysis of the Sample (AS)	1. Calculate the appropriate test statistic so that it addresses the research question. 2. Determine the appropriate sampling distribution (the distribution of the statistic conditioned on the null hypothesis). 3. Use the sampling distribution of the statistic and the decision rule to determine whether or not the sample is unusual under the null condition. If so, the results are said to be statistically significant.	• For a given sample size, *n*, and sample statistic, the sampling distribution of the statistic gives a probability distribution of values taken by the sample statistic for all possible samples of size *n*. [Note: It does not give the distribution of values for a particular sample.] • In order to determine if the sample is unusual under the null condition, one should examine the sampling distribution of the given statistic, for samples of size *n* that describes the distribution if the null hypothesis described the true nature of the population. • Conditioned on the null, the sampling distribution of the statistic for samples of size *n* gives the probability (*p*-value) of getting values of the test statistic at least as extreme as the observed value, if the null were true. • If the probability is small, and statistical significance is achieved (the *p*-value is less than or equal to α), then the null is likely not true. If not, then there is support for the null and the results are not statistically significant.
Conclusion (C)	1. Make a decision as to whether to *retain* or *reject* the null hypothesis. This decision should be based on analysis of whether or not the sample is unusual (statistical significance is achieved) under the null condition. 2. State the conclusion to retain or reject the null hypothesis and indicate that conclusion has not been proven; rather it is the result of an inference about a population	• The logic of proof provides the foundation for statistical hypothesis testing o Generation of supporting examples does not prove and, thus, a decision to retain the null means only that there is some support for the null situation. It is not a proof. o Generation of one counter-example can

	using information from a sample, probabilistic reasoning, and logic. 3. Interpret the conclusion to reject the null as an indication of the degree to which the sample represents a "counterexample" to the null hypothesis. .	disprove an assumed hypothesis but determining whether or not a sample is impossible under the assumed null hypothesis is seldom possible for large populations. Thus, any decision to reject the null indicates that the sample is unusual by some probabilistic standard. It is unlikely under the null and, since the alternative is a contradictory descriptor, more confidence is placed in the alternative as the true descriptor. • The population for which the conclusion applies is dependent upon the sample collected. • The sample statistics collected describe the samples themselves. These statistics are used to draw inferences about hypotheses that address the corresponding population parameters, using probabilistic and logical reasoning. The sample statistics cannot, therefore, be directly converted to population parameters.
Implication for Practice (IP)	1. Recognize the tension between statistical and practical significance. 2. Recognize the need to consider sample size and effect size (at least informally) when deciding whether or not statistically significant differences do/do not indicate real or practical differences in the population. 3. Consider implications for Type I and Type II error as well as practical considerations when making a decision for action based on results from a statistical hypothesis test.	• Statistical hypothesis testing is not a proof and this should be taken into consideration when making decisions based on the results • The implications for error (Type 1 and Type II) must be taken into consideration when making a decision based on the results. • Finding that there are no statistically significant differences between two groups does not necessarily mean that the populations aren't different – especially when the sample size is small. • Although the difference between two groups is found to be statistically significant, one should investigate the actual difference before making any decisions based on these results

Each major section of the framework presents processes involved in one key stage of a statistical hypothesis testing study and an explanation of the understandings that are fundamental to work on that task. For convenience of reference, each process is labeled with a mnemonic code. For example, **RA** 1.1, 1.2, 1.3 and **RA** 2 identify the tasks of generating competing hypotheses in the stage of study design where one recognizes that hypothesis testing would be an appropriate research approach to a question.

Items used on the multiple-choice instrument address the categories outlined in the Framework for Assessing Understanding. The construction of the various distractors for the multiple-choice items took into account both the understanding(s) to be assessed and the informal ways of reasoning identified in the literature review. The result is a collection of distractors that, if chosen, would indicate that either (a) the individual does not possess the understandings being assessed in that item or (b) the individual is reasoning according to identified informal heuristics of misconception. See Appendix A for a list of the final multiple-choice items and accompanying justifications for the distractors associated with each item.

Because the goal of this study was to explore and report on the nature of student understanding, data analysis was primarily descriptive in nature. Overall scores earned by students and descriptive, summary statistics such as mean, median, upper and lower quartiles, and range were calculated on both assessments only for those who completed both. Performance on individual items was calculated, including the frequency with which various distractors were chosen. In addition, aggregate scores for framework categories were calculated. Analysis of these results provides information about how introductory statistics students understand statistical hypothesis testing.

Further analysis with respect to the research sub-questions was performed. To address research sub-question number one, the summary statistics of the course exam and the multiple-choice assessment were compared and a correlation analysis was performed to determine whether the scores correlate.

To address research sub-questions two and three, each item on the multiple-choice assessment was classified according to whether it assesses (1) understanding of the logic and reasoning of statistical hypothesis testing or (2) understanding of the relationship between the statistical hypothesis testing method and the context in which it is employed. These categories cross over the Framework categories. For example, in addition to assessing whether an individual recognizes when to use statistical hypothesis testing, an item in the **RA** category might also assess whether an individual understands either (a) the ***logic and reasoning*** of statistical hypothesis testing or (b) the relationship of the ***method and context***. Table 3.2 illustrates the double classification for each item.

Table 3.2

Item Classification, Multiple-Choice Assessment

Item Number	Framework Category	Research Sub-question Addressed
1	RA	Method and Context
2	RA	Method and Context
3	RA	Logic and Reasoning
4	GH	Logic and Reasoning
5	C	Logic and Reasoning
6	C	Logic and Reasoning
7	GH	Logic and Reasoning
8	AS	Logic and Reasoning
9	AS	Logic and Reasoning
10	CS	Method and Context
11	DR	Logic and Reasoning
12	IP	Method and Context
13	AS	Logic and Reasoning
14	IP	Method and Context

Overall percentage correct was calculated for the ***logic and reasoning*** items as well as for the ***method and context*** items. These analyses were enhanced by individual item analyses and by analyses of groups of items that represent various framework categories. Combined, the various analyses provide a more complete description of student understanding with respect to the research question and sub-questions.

Qualitative Phase: Data and Analysis

In order to provide more insight into student thinking, follow-up interviews were conducted with a group of students who represented a range of performance patterns on the two quantitative assessments. Given summary statistics for the two assessments, participants were put into one of four categories based on the quartile placement of their scores on the two assessment instruments. These categories are illustrated in Table 3.3.

Table 3.3

Follow-Up Interview Participant Classification

Category	Course Exam Quartile	Multiple-choice Assessment Quartile
HH	Top (High)	Top (High)
HL	Top (High)	Lower (Low)
LH	Bottom (Low)	Top (High)
LL	Bottom (Low)	Bottom (Low)

An individual in the **HL** category, for example, scored in the top quartile on the course exam but scored in the bottom quartile on the multiple-choice assessment. Given the groups, three individuals were randomly chosen from the **HH** group, four from the **HL** group, and four from the **LH** group and asked to participate in the follow-up interview. In choosing students who scored in the top and bottom quartiles, it was possible to gain insight into student thinking representative of the various performance patterns on the two quantitative assessments. In addition, this selection process maximized the potential for variability of responses. Note that more participants were chosen from the **HL** and **LH** groups than from the **HH** group. One of the reasons for conducting this study is a belief (generally as a result of anecdotal evidence) that, although students may be able to perform (some of) the steps involved in a well defined statistical hypothesis testing problem, it does not necessarily mean that they have well-developed conceptual understandings of the method and its uses. Participants from these two groups can give insight into the why this may or may not be the case. Individuals from the **LL** group were not chosen to participate because students who scored in the bottom quartiles would not be useful in addressing the research question and sub-questions.

Using the results of the multiple-choice assessment, interview items were chosen and students were asked to explain their thinking on each. From the pilot study, it was determined that only 9 of the 14 items could be discussed in the hour-long interview. In order to answer the second and third research sub-questions, it was decided that the interview should address half of the ***logic* and *reasoning*** items (six items) and half of the ***method* and *context*** items (three items). These items were chosen so that they: (1) addressed as many categories of the framework as possible; and (2) were items for which there was low performance. This approach to eliminating items from the assessment to be included in the interview allowed for the research sub-questions to be addressed. It also provided a means for gaining insight into the various ways that students think about and understand statistical hypothesis testing.

Students were provided with their multiple-choice assessment, a pencil, and paper. Every attempt was made to ensure the interview was conducted as a natural conversation between interviewer and interviewee. In order to situate that interview, participants were first asked to describe statistical hypothesis testing and explain how it could be used. After some discussion about the method and its uses, the interview continued with discussion of the identified items. The interviewees were asked to explain why they chose the answer they chose to each of the nine identified items and why they didn't choose the other answers. Participants were asked follow-up probing questions to further clarify their thinking. If students cited "rules" learned in class, they were asked if these "rules" made sense to them. In addition, prior to conducting the interviews, an analysis of the item and its relation to the research question combined with the results from the quantitative phase provided a list of issues for which it was important to address in each item. If these issues were not naturally addressed by the participant in his/her explanations, they were raised as additional follow-up questions.

The interviews were audio-taped, transcribed, and analyzed for commonality as well as uniqueness among answers. Initial analysis of the data was conducted *within* groups (**HL**, **LH**, or **HH**) for each item included in the follow-up interview. This analysis of student thinking was subsequently conducted *across* the groups of participants for each item. Ultimately, analysis of student thinking was conducted within categories of items that addressed the various research sub-questions. Summaries of commonality and uniqueness among student responses were constructed and reported.

The information obtained from the interviews extends the information obtained from responses to the multiple-choice instrument, and also provides a form of validity check for that instrument. As was described above, the multiple-choice assessment places limitations on the ways that students can respond to the items. In constructing the questions and possible choices, assumptions were made about the potential ways that students might think. The answer options provided may not match the way that an individual student is thinking. The interview provides a test of validity on those assumptions. Overall, the interview also tested the validity of the claims about student understanding that might be made based solely on results obtained from the multiple-choice instrument. It is through this triangulation of data sources that conclusions are better substantiated (Greene, 2001).

Summary

The mixed methods approach used in this study is an effective means of collecting information about student understanding of statistical hypothesis testing for a large number of students, while also gaining more insight into student thinking than can be attained from quantitative measures alone. Each phase of the study (quantitative and qualitative) was useful in addressing the research question and sub-questions and provided valuable information that fills a gap in the literature. The results of the study are outlined in the next two chapters.

CHAPTER 4

QUANTITATIVE RESULTS AND ANALYSIS

In this chapter, the results of the quantitative phase the study will be presented and analyzed. A total of 104 students from 8 sections of STAT 100 participated in this phase of the study, completing both the course exam and the multiple-choice assessment. Analysis of the results yielded information that addresses the three identified research sub-questions and provides large scale descriptive information about student understanding of statistical hypothesis testing, overall. The chapter is organized into three sections: (1) presentation of the results, (2) analysis of the data, and (3) summary of the data and conclusions associated with the quantitative phase of the study.

Results

Course Exam

The third course exam was given to the participants during class time, was standard across all eight sections (with several similar forms), and was graded by the course instructors. The questions were written to assess student understanding of statistical hypothesis testing and, as demonstrated in Chapter 3, were largely measures of procedural knowledge. The questions were "free response" in that students were asked to solve problems and to show their work. Students were provided with statistical tables and formula sheets. Calculators were permitted. The exam was scored out of 100 possible points.

Descriptive, summary statistics associated with the results of the course exam for those students participating in the study only (*n* = 104) are reported in Table 4.1.

Table 4.1

Descriptive Statistics, Course Exam, *n* = 104

Summary Statistic	Value
Mean	73.31
Standard Deviation	19.906
Minimum Score	16
Maximum Score	99
Quartile 1	59
Quartile 2 (Median)	79.5
Quartile 3	90

Overall, students performed well on the exam. The average score was a 73.31, with a median score of 79.5 out of 100 points possible. Though the range of scores runs from 16 to 99, 75% of these scores were above 60 points (as indicated by the first quartile score). Half of the students earned at least 79.5 points (as indicated by the median).

It should be noted this sample of students is representative of the overall population of students enrolled in STAT 100 at the time of the study. Descriptive, summary statistics for overall student performance on the third course exam (*n* = 218) are reported in Table 4.2.

Table 4.2

Descriptive Statistics, Course Exam, *n* = 218

Summary Statistic	Value
Mean	71.11
Standard Deviation	19.72
Minimum Score	10
Maximum Score	100
Quartile 1	61.88
Quartile 2 (Median)	76
Quartile 3	90.13

In comparing Tables 4.1 and 4.2, we see that, with the exception of the minimum score, summary statistics for the overall performance of the population of students enrolled in STAT 100 differ from those of the sample by no more than 3.5 points. The minimum scores differ by only 6 points. This comparison indicates that, with regard to performance on the third course exam, the sample of students who participated in the study is representative of the overall population of students enrolled in STAT 100 at the time of data collection in the quantitative phase.

Multiple-Choice Assessment

The multiple-choice assessment was administered to participating students during class time. Students were not provided with statistical tables nor were they provided with formula sheets as these items were not needed to answer the questions found on the assessment. In order to limit the number of contexts participants would be exposed to as they read the items on the assessment, the items were organized around common, contextual themes: study of educational program effectiveness, study of product quality, and study of student traits. The assessment was scored out of 14 points, one point per question. The multiple-choice assessment is included in Appendix B.

Descriptive, summary statistics associated with the results of the multiple-choice assessment are reported in Table 4.3.

Table 4.3

Descriptive Statistics, Multiple-Choice Assessment, *n* = 104

Summary Statistic	Value (percent correct)
Mean	4.53 (32.4%)
Standard Deviation	1.87
Minimum Score	0
Maximum Score	10 (71%)
Quartile 1	3 (21.4%)
Quartile 2 (Median)	4 (28.6%)
Quartile 3	6 (42.9%)

Performance on this assessment was lower than that of the course exam. The average score was 4.53 points out of 14, roughly 32.4% of the total points. The scores ranged from 0 to 10 points, with only one participant earning 10 points. Seventy-five percent of the participants scored no more than 6 points (42.8% of 14 possible points), and only half of the students scored above 4 points (28.6% of 14 possible points).

Performance on individual items is given in Table 4.4. For each item, Table 4.4 reports the number and percentage of participants who chose the correct answer. These values are shown in **bold** print. In addition, if 20% or more of the participants chose a given distractor, the frequencies and percentages associated with those distractors are reported.

Table 4.4

Frequency of Responses, Multiple-Choice Assessment

Item Number	Category	Percent Correct	Answer Choice	Frequency	Percent of Respondents
1	Recognizing Applicability	43.27	**a**	**45**	**43.4**
			d	54	51.9
2	Recognizing Applicability	61.54	b	26	25.0
			d	**64**	**61.5**
3	Recognizing Applicability	11.54	**a**	**12**	**11.5**
			b	37	35.6
			c	25	24.0
			d	30	28.8
4	Generating Statistical Hypotheses	30.77	a	51	49.0
			c	**32**	**30.8**
5	Conclusion	6.73	c	91	87.5
			d	**7**	**6.7**
6	Conclusion	36.54	b	36	34.6
			c	**38**	**36.5**
			d	22	21.2
7	Generating Statistical Hypotheses	51.92	a	21	20.2
			d	**54**	**51.9**
8	Analysis of the Sample	33.65	**a**	**35**	**33.7**
			b	53	51.0
9	Analysis of the Sample	40.38	**a**	**42**	**40.4**
			b	22	21.2
			d	29	27.9
10	Collect a Sample	45.19	b	34	32.7
			c	**47**	**45.2**
11	Decision Rule	29.81	a	35	33.7
			b	30	28.8
			c	**31**	**29.8**
12	Implication for Practice	36.54	**a**	**38**	**36.5**
			b	21	20.2
			d	25	24.0
13	Analysis of the Sample	11.54	a	31	29.8
			c	47	45.2
			d	**12**	**11.5**
14	Implication for Practice	13.46	a	34	32.7
			b	26	25.0
			c	29	27.9
			d	**14**	**13.5**

As is evident from Table 4.4, student performance on individual items was low, ranging from 6.73% to 61.5% correct on a given item. Student performance was strongest on item number 2. This item (classified in the **Recognizing Applicability** category) assesses whether students understand that statistical hypothesis testing is only useful in answering research questions when the question can be answered by some *measure* of the population. Relatively speaking, participants also did well on item number 7 (classified in the **Generating Statistical Hypotheses** category), which assessed whether students could choose, from the listed possibilities, the correct alternative and null hypotheses to match the given situation.

Student performance was weakest on item number 5. This item (classified in the **Conclusion** category) assesses whether, given information about statistical significance, students are able to (1) draw the correct conclusion about the null hypothesis and (2) interpret that conclusion as an indication of the degree to which the sample represents a "counterexample" to the null hypothesis. Relatively speaking, student performance was also low on item numbers 3 and 13. Item number 3 (classified in the **Recognizing Applicability** category) assesses whether students understand statistical hypothesis testing as a means not only to answer a research question, but also to quantify the uncertainty associated with that answer. Item number 13 (classified in the **Analysis of the Sample** category) assesses student understanding of *p*-value as a measure of unusualness of the sample under the assumed null condition.

It is interesting to note that, for some items, most participants chose between only *two* possible answers, while for others, participants chose from among all four answer choices. Table 4.4 illustrates that variation in answers for each of the 14 items. For example, in looking at item number 5, we see that 91 of 104 participants chose option *c* while only 7 of the participants chose the correct answer, *d*. However, in looking at item number 3, we see that all four answer choices were relatively popular. This additional information is of interest. In both cases, performance was low. However, the reason performance was low in number 5 is different than for number 3. This difference indicates that, with regard to the issue addressed in number 5, students complete introductory statistics courses with one clear misunderstanding. Whereas, with regard to the issue addressed in number 3, students have a variety of misunderstandings. Similar analyses can be performed with other items and is useful to analyze the data with respect to the research sub-questions identified for this study.

Aggregation of the data with respect to the categories identified by the Framework for Assessing Understanding provides information about whether students do or do not have the desired understandings of the process identified in the Framework. Table 4.5 gives aggregate results for each category of the Framework for Assessing Understanding.

Table 4.5

Multiple-Choice Results by Framework Category

Category	Item	Percent Correct	Average
Recognizing Applicability	1	.4327	
(RA)	2	.6154	.3878
	3	.1154	
Generating Statistical Hypotheses	4	.3077	.4135
(GH)	7	.5192	
Decision Rule	11	.2981	.2981
(DR)			
Collect a Sample	10	.4519	.4591
(CS)			
Analysis of the Sample	8	.3365	
(AS)	9	.4038	.2852
	13	.1154	
Conclusion	5	.0673	.2164
(C)	6	.3654	
Implication for Practice	12	.3654	.25
(IP)	14	.1346	

Given the percent correct for each item, the column labeled "Average" indicates the percent of items answered correctly in a given *category*. For example, the table tells us that 38.78% of the items in the **RA** category were answered correctly [(.4237 + .6254 + .1154) / 3 = .3878].

In looking at the "Average" column in Table 4.5, we see that the participants scored relatively well on the **Generating Statistical Hypotheses** items with 41.35% of the items answered correctly. In addition, participants scored well in the **Collect a Sample** category. However that category was represented by only one item. Performance in the **Recognizing Applicability** category is also relatively high. Participants did not score well in the **Analysis of the Sample**, **Conclusion**, and **Implication for Practice** categories with 28.52%, 21.64%, and 25% of the students choosing correct answers, respectively. Based on this analysis, it seems that introductory students are better able to generate hypotheses then they are to analyze the sample, to draw a conclusion, and to interpret that conclusion for use in practice.

The results presented in this section provide information on student understanding as measured by the two assessment instruments. In the next section, analysis of results within and across the instruments will be presented with respect to the research question and sub-questions identified in this study.

Analysis

This study was an exploration of student understanding of the "big picture" of statistical hypothesis testing and it addressed the following overarching research question: *What are the understandings of statistical hypothesis testing held by students who have completed a traditional, introductory course in statistics at a large university?* Given the conceptual analysis and theoretical perspective for understanding outlined in Chapter 1, three related research sub-questions were identified:

1. *What is the relationship between introductory students' understanding of the procedures and the concepts, logic, and uses of statistical hypothesis testing?*
2. *What are the understandings that introductory students have of the overall logic and reasoning of statistical hypothesis testing?*
3. *What are introductory students' understandings of the relationship between the method of statistical hypothesis testing and the context in which it is employed?*

Analysis of the data both at the aggregated and individual levels provides information useful in answering these three research sub-questions, and it provides insight into the general, overarching research question. Each of the research sub-questions will be considered in turn.

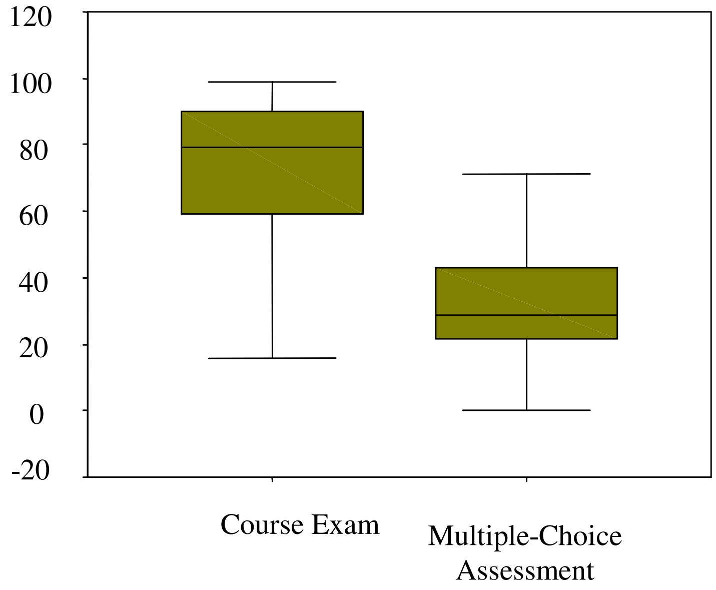
Research Sub-question Number 1

Comparison of the results on the two quantitative assessments provides data that is useful in describing the relationship of students' understanding of the procedures and the concepts, logic, and uses of statistical hypothesis testing. In comparing the two, we see that the participants scored very differently on the two assessments, which indicates that there is not a strong relationship between student understanding of the hypothesis testing procedure and his/her understanding of the supporting logic, concepts, and uses.

Referring to the descriptive, summary statistics found in Tables 4.1 and 4.3 we see that, on all summary measures, the multiple-choice assessment ranks lower than the course exam. In addition, the spread of scores is different on the two assessments. Box plots for each assessment illustrate the stark differences in the spread of scores between the two assessments.

Figure 4.1

Box Plot, Quantitative Assessment Scores, *n* = 104



Histograms that represent the distribution of scores are provided in Figures 4.2 and 4.3. Here we see differences in the way the data is “clumped” on each assessment.

Figure 4.2

Histogram, Scores Course Exam, *n* = 104

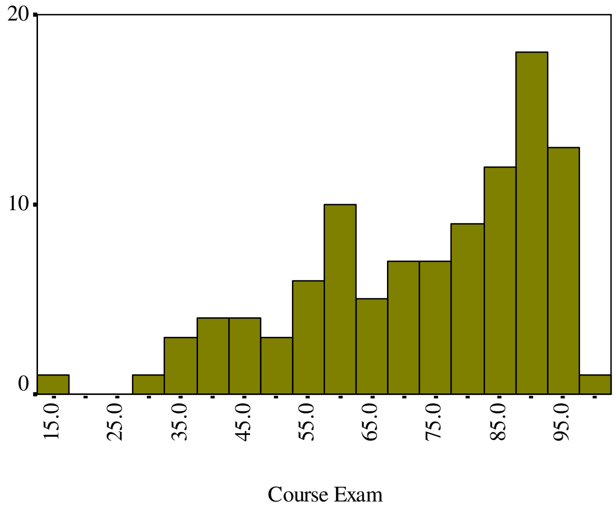
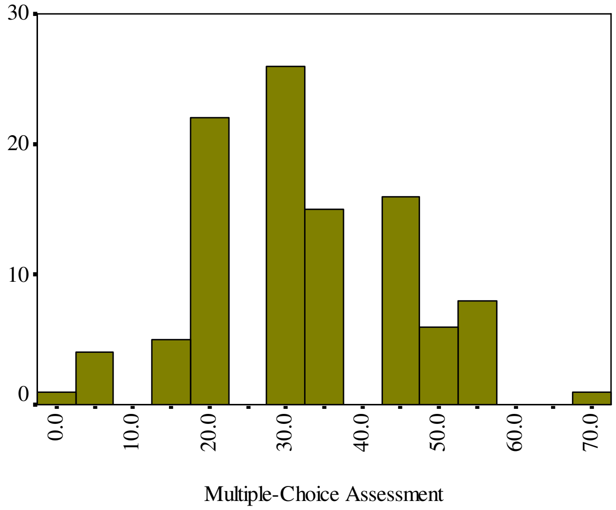


Figure 4.3

Histogram, Scores Multiple-Choice Assessment, *n* = 104

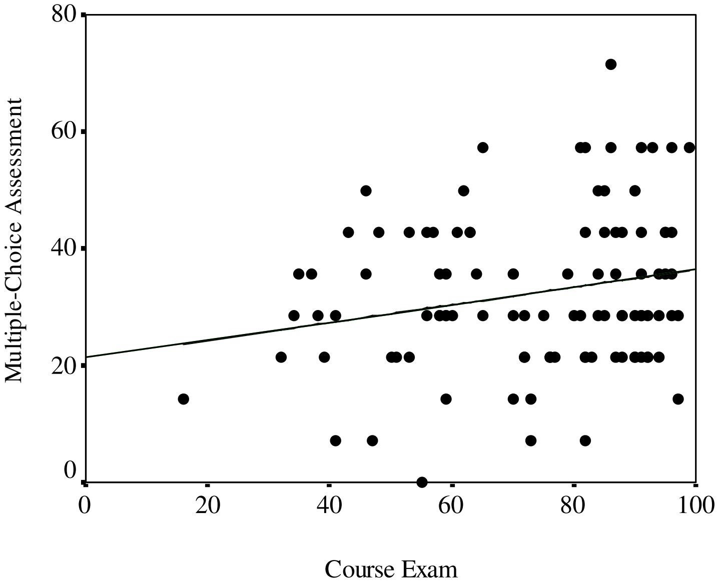


The strong performance on the course exam was expected. As was mentioned in the first chapter, introductory statistics courses tend to focus instruction on student ability to employ the procedure for statistical hypothesis testing (Garfield, Hogg, Schau, & Whittinghill, 2002; Shaughnessy, 1992). If this is the case, students should do well on assessments that focus on *procedural fluency*. That was the case for the participants in this study. Unfortunately, though the participants performed well on the course exam, they did not perform well on the multiple-choice assessment.

Further analysis indicates not only that, on the whole, scores on the course exam were lower than on the multiple-choice assessment, but also that scores on the two assessments were not strongly related. That is, students who scored well on the course exam did not *necessarily* score well on the multiple-choice assessment and vice-versa. A scatter plot of scores for each student (*n* = 104) illustrates this phenomenon.

Figure 4.4

Scatter Plot and Regression Line, Quantitative Assessments, *n* = 104



The regression line included in Figure 4.4 indicates a weak correlation between the two quantitative scores. In fact, analysis of the correlation between scores on the course exam and the multiple-choice assessment confirms that the strength of the relationship between the two is weak.

Table 4.6

Pearson Correlation, Quantitative Assessments

Pearson Correlation	.215
p-value	.028

Table 4.6 indicates that the Pearson correlation between the two assessments is .215. While this figure is significantly different from 0 at the $\alpha = 0.05$ level, it is not a strong correlation. The weak correlation indicates that, a student's score on the course exam is not a strong predictor of his/her score on the multiple-choice survey. That is, mastery of the hypothesis testing procedure does not necessarily mean strong understandings of the concepts, logic, and uses of the procedure and vice versa.

Overall, the results indicate that the relationship between student understanding of the procedures and of the concepts, logic, and uses of statistical hypothesis testing is weak. In as much as the course exam is a measure of *procedural fluency* and the multiple-choice assessment a measure of *conceptual understanding*, *adaptive reasoning*, *strategic competence*, and *productive disposition* (with respect to understanding the *value* of the method), we see that strong performance on assessments focused on application of the procedure to well-defined, traditional problems does not give a measure (even relatively speaking) of the degree to which students have a deep, connected understanding of that procedure and its uses. On the whole, students in a fairly typical introductory statistics course demonstrated strong understandings of the hypothesis testing procedure and weak understandings its logic, concepts and uses. Examination of this data with respect to the other two research sub-questions gives more insight into why student performance on the multiple-choice assessment was low.

Research Sub-question Number 2

Analysis of the results associated with the items classified as ***logic and reasoning*** is useful to better describe how introductory students understand the overall logic and reasoning of statistical hypothesis testing. Table 4.7 gives aggregate data for items in this category.

Table 4.7

Logic and Reasoning Category, Multiple-Choice Results

Item	Framework Category	Percent Correct	Average
3	Recognizing Applicability	11.54	
4	Generating Statistical Hypotheses	30.77	
5	Conclusion	6.73	
6	Conclusion	36.54	28.10
7	Generating Statistical Hypotheses	51.92	
8	Analysis of the Sample	33.65	
9	Analysis of the Sample	40.38	
11	Decision Rule	29.81	
13	Analysis of the Sample	11.54	

The average percent correct on these 11 items was 28.10%. However, there are vast differences in the percentage of students who correctly answered individual items. A relatively strong performance was reported for item number 7 in the **Generating Statistical Hypotheses** category. Relatively low performances were reported for item numbers 8 and 13 in the **Analysis of the Sample** category as well as for item numbers 5 and 6, both from the **Conclusion** category. Both the **Decision Rule** and **Recognizing**

Applicability categories were represented by only one item and performance on those items was also relatively low.

The difference between scores on the **Generating Statistical Hypotheses** and other categories makes sense if we consider the items themselves. As demonstrated in the previous section, students generally did well on the course exam, a measure of *procedural fluency*. Item numbers 4 and 7 ask students to state the hypotheses associated with the situation presented in the stem. This process can become very procedural. If one merely associates the null hypothesis with a statement of equality, then he/she is able to narrow down the possible options provided in each item. This elimination of answer choices increases the chance of choosing the correct answer. However, in order to choose correctly from among the remaining answer choices, one must recognize that the alternative and the null hypotheses must be contradictory. Relative success on items 4 and 7 might suggest that students have some understanding, but that procedural knowledge may have played a major role.

The items contained in the other categories can not necessarily be answered through knowledge of the procedures. Analysis at the level of individual items within the **Analysis of the Sample**, **Decision Rule**, **Conclusion**, and **Recognizing Applicability** categories gives insight into how students understand the issues addressed by each of these categories. This analysis was used to identify several areas of difficulty and/or misunderstandings. Ultimately, this analysis provides information useful to describing student understanding of the logic and reasoning of statistical hypothesis testing. The analysis is broken into two parts: analysis of items representing the **Analysis of the**

Sample and **Decision Rule** categories and analysis of items representing the **Conclusion** and **Recognizing Applicability** categories.

Analysis of the Sample and Decision Rule Categories

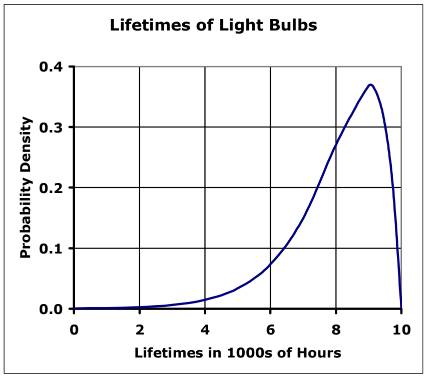
The **Analysis of the Sample** and **Decision Rule** categories assess whether students understand the role of sampling distributions and probability in the logic and reasoning of statistical hypothesis testing. Sampling distributions are used to determine the probability of samples at least as extreme as the observed sample, under the null condition. If this probability is small, the results are statistically significant and the null is rejected. If not, the null is retained. The significance level gives the "cut point" for a small probability and is the probability of a Type I error if the null is, indeed, correct. Analysis of the items within these categories indicates that students do not have a strong understanding of the statistical concepts involved here, nor do they understand the overall reasoning involved. Three themes of student understanding emerged from this analysis and they will be discussed in turn.

Conceptual Understanding of Sampling Distribution. Analysis of responses to item numbers 8 and 9 provides evidence that students do not have strong understandings of sampling distributions and how they are used in statistical hypothesis testing. Both items are classified in the Analysis of the Sample category and assess understanding of the logic and reasoning of statistical hypothesis testing. In order to answer the items correctly, though, one must understand sampling distributions and their role in the statistical hypothesis testing. Consider item number 8.

Figure 4.5

Item Number 8, Multiple-Choice Assessment

8. The typical distribution of usable lifetimes for light bulbs is shown in the following sketch. It's clearly not a normal distribution. However, when doing a test to compare the mean lifetimes of two light bulb brands (using large samples of both brands), a statistician used a normal distribution to find the *p*-value of the difference.



Which of the following statements ***best explains why*** the test for significance involves normal probabilities, rather than probabilities from the light bulb lifetime distribution?

a. The distribution of the difference of means of large samples is always approximately normal.

b. The distribution of values in large samples is always approximately normal.

c. Values of the standard normal probability distribution are always given in reference tables.

d. The distribution of differences of values in two samples is always approximately normal.

This item assesses student understanding of sampling distributions as useful to reasoning applied in statistical hypothesis testing. In order to answer the question correctly students must understand that sampling distributions can be used to find probabilities associated with obtaining a range of sample statistics, for a given sample size, if the population parameter is known (or assumed to be a specific value). They must also understand that, for tests of differences of means, the distribution of differences of sample means is normal. The probability of the observed result or one more extreme (under the null condition) is the *p*-value. If the *p*-value is small enough, the null hypothesis is rejected.

If students choose the correct answer, *a*, they demonstrate an understanding of sampling distribution as distribution of values for many samples used to give probabilities conditioned on the null. In this case, that distribution of values for many samples is a normal distribution. If a student chooses answer choice *b* or *d*, this would indicate that the student does not understand what a sampling distribution represents. If a student chooses answer choice *c* over the other options, this would indicate that he/she does not have an understanding of the role of the normal distribution in the process.

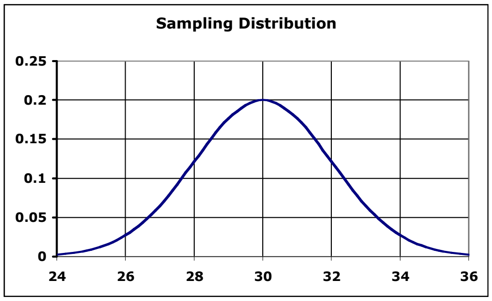
Student performance on this item was not strong. As reported in Table 4.4, only 33.65% of the students chose the correct answer. However, 51% chose incorrect option *b*. This indicates that students do not understand that *sampling distributions* of means (or differences of means) are normally distributed and that they give the distribution of differences of means for *many* samples. Rather, students think that the values in the sample *itself* are normally distributed.

The results associated with item number 9 give further evidence that students do not have a strong understanding of sampling distributions and their role in statistical hypothesis testing.

Figure 4.6

Item Number 9, Multiple-Choice Assessment

9. Tests show that fuel efficiency of cars in the current model year averages 30 miles per gallon. A test of 100 new car models gave mean fuel efficiency of 31.5 miles per gallon. To see whether it is correct to claim that the new car models are more fuel-efficient than those in the current model year, a researcher constructed the sampling distribution of average fuel efficiencies of many samples of 100 current model cars shown in the following graph.



Which of these conclusions is ***best supported*** by the graph above?

a. The difference in fuel efficiency of current and new model cars is not statistically significant.

b. Half of current and new model cars have fuel-efficiencies below 30 miles per gallon.

c. The difference in fuel efficiency of current and new model cars is statistically significant.

d. Nothing can be concluded. The graph should be centered at 31.5.

In order to answer the question correctly, students should understand that because the sampling distribution is a distribution of values of the test statistic for all possible samples of a given size conditioned on the null hypothesis, it gives probabilities useful in determining whether the observed result is unusual under the null condition. Additionally, these probabilities are given by areas under the curve. If the probability of the observed result or more extreme is small, the results are said to be statistically significant and the null hypothesis is rejected. If not, the results are not statistically significant and we fail to reject the null hypothesis.

Answer choice *a* is correct and answer choice *c* states the opposite. If a student chooses answer choice *b*, this would indicate that the student does not understand what a sampling distribution represents. They have a weak conceptual understanding of sampling distributions. Answer choice *d* was written to assess whether students understand that the test for unusualness is conditioned on the null. However, it could be argued that students who do not have a conceptual understanding of sampling distributions might not understand what the graph represents and, therefore, choose *d*.

According to Table 4.4, 40.4% of the participants chose the correct answer, *a*. However, 27.88% chose *d* and 21.11% of the participants chose *b*. Such results provide evidence that students may not understand the concept of sampling distributions and, therefore, do not understand the overall *logic and reasoning* associated with statistical hypothesis testing.

Combined, the results associated with item numbers 8 and 9 indicate that students do not have strong understandings of sampling distributions and their role in statistical hypothesis testing. They seem to struggle with the concept of sampling distribution as a distribution of sample statistics of *all possible samples* and, instead, understand them to represent the distribution of values *within* a sample.

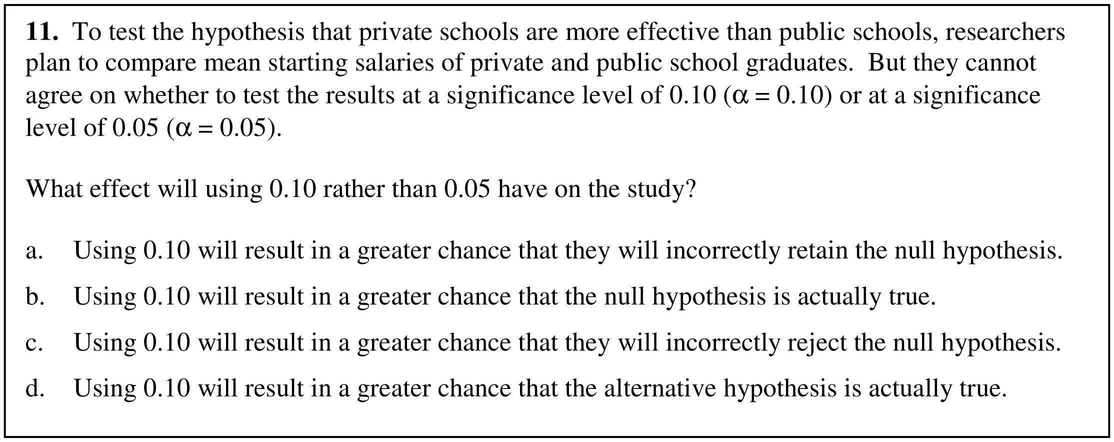
Conceptual Understanding of Significance. Related to the role of sampling distributions in the logic and reasoning of statistical hypothesis testing, is that of significance level. The sampling distribution and significance level are both referenced when determining whether the sample is unusual, conditioned on the null hypothesis. In practice the decision is made in one of two ways: (1) by using sampling distributions and the level of significance, α, to establish a critical value and by comparing the test statistic to that value; or (2) by using statistical tables to determine the *p*-value associated with the data and comparing that to the significance level, α. In both cases, it is necessary to refer to that level of significance, α, to make a decision whether or not to reject the null hypothesis.

As was described in the Conceptual Analysis presented in Chapter 1, the level of significance (α) used in a statistical hypothesis test plays an important role in determining the degree of certainty one has with his/her decision to reject the null hypothesis. It is the probability of making a Type I error, if the null is, indeed, true. That is, it is the probability that an individual will *incorrectly* reject the null hypothesis. The greater the level of significance used, the greater the probability of rejecting the null hypothesis under the condition that the null hypothesis describes the true nature of the situation. The level of significance provides a "cut point" with which to determine whether or not a sample is unusual under the null condition. It is a probability used to determine if the probability of the observed statistic is small enough to deem the sample an unusual (unexpected) occurrence if the null hypothesis is, indeed, correct. If the sample is unlikely, then less confidence is place in the null hypothesis and it is rejected.

Item number 11 assesses whether or not students have this understanding of significance level. It is classified in the **Decision Rule** category and assesses student understanding of significance level in the overall ***logic and reasoning*** of statistical hypothesis testing.

Figure 4.7

Item Number 11, Multiple-Choice Assessment



11. To test the hypothesis that private schools are more effective than public schools, researchers plan to compare mean starting salaries of private and public school graduates. But they cannot agree on whether to test the results at a significance level of 0.10 ($\alpha = 0.10$) or at a significance level of 0.05 ($\alpha = 0.05$).

What effect will using 0.10 rather than 0.05 have on the study?

a. Using 0.10 will result in a greater chance that they will incorrectly retain the null hypothesis.

b. Using 0.10 will result in a greater chance that the null hypothesis is actually true.

c. Using 0.10 will result in a greater chance that they will incorrectly reject the null hypothesis.

d. Using 0.10 will result in a greater chance that the alternative hypothesis is actually true.

Answer choice *c* is correct. Answer choices *b* and *d* were written to assess whether or not a student understands the level of significance as the probability that the null hypothesis is actually true or untrue, rather than the probability of committing an error, if the null hypothesis is true. Answer choice *a* is the opposite of *c* and will indicate whether or not students understand that a larger significance level gives more chance that the null hypothesis is rejected, rather than retained.

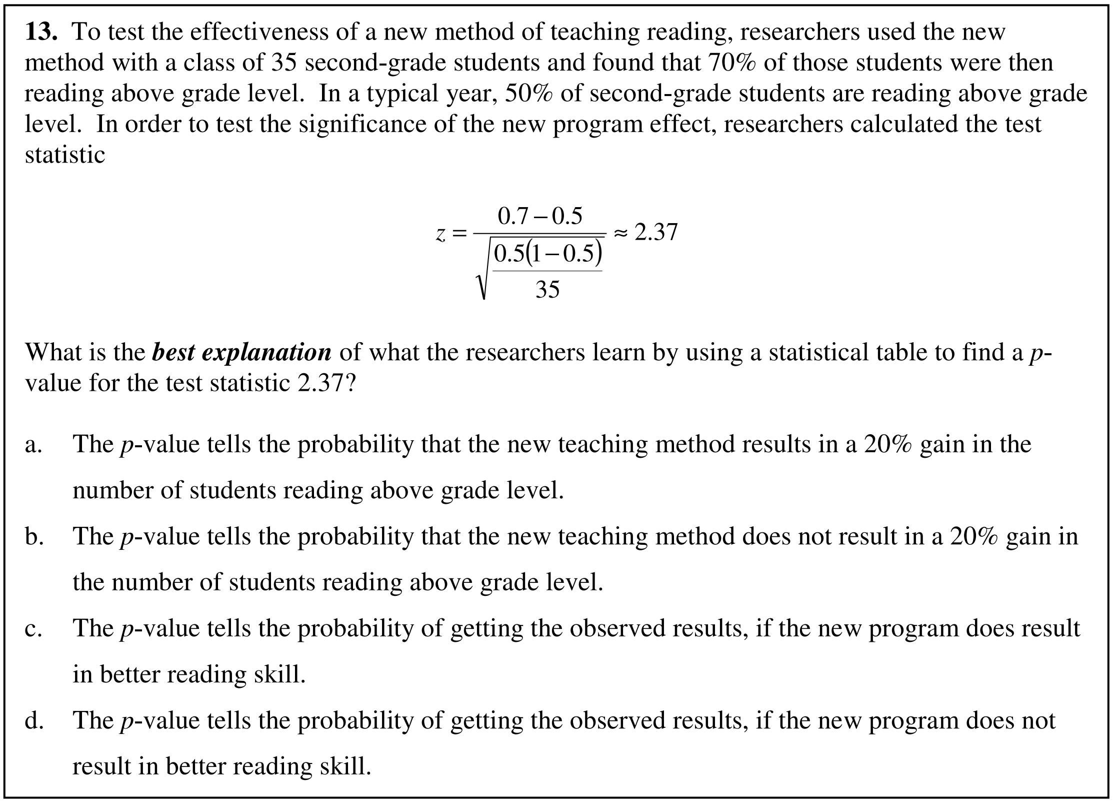
Student performance on this item was low. Only 29.8% of the participants chose answer choice *c*. Options *a* and *b* were popular, with 33.7% of the participants choosing *a* and 29.8% of the participants choosing *b*. These results indicate that students do not have a strong understanding of the significance level and its role in statistical hypothesis testing.

Conceptual Understanding of p-value. As mentioned above, the *p*-value often plays a role in the analysis of a sample. The *p*-value is the probability that the observed value, or more extreme, would occur if the null hypothesis described the true nature of the situation. It is a value that is attached to the sample. This value is useful in statistical hypothesis testing in that it gives a measure of "unusualness" for the sample, conditioned on the null. If the *p*-value associated with a sample is small, the sample is considered to be an unlikely occurrence if the null were true. Since the sample was randomly chosen from the population, the production of an unusual sample leads to less confidence that the null is correct.

Item number 13 on the multiple-choice assessment asks students to choose the option that correctly "explains" what the *p*-value represents and how it is used in statistical hypothesis testing.

Figure 4.8

Item Number 13, Multiple-Choice Assessment



13. To test the effectiveness of a new method of teaching reading, researchers used the new method with a class of 35 second-grade students and found that 70% of those students were then reading above grade level. In a typical year, 50% of second-grade students are reading above grade level. In order to test the significance of the new program effect, researchers calculated the test statistic

$$z = \frac{0.7 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{35}}} \approx 2.37$$

What is the ***best explanation*** of what the researchers learn by using a statistical table to find a *p*-value for the test statistic 2.37?

a. The *p*-value tells the probability that the new teaching method results in a 20% gain in the number of students reading above grade level.

b. The *p*-value tells the probability that the new teaching method does not result in a 20% gain in the number of students reading above grade level.

c. The *p*-value tells the probability of getting the observed results, if the new program does result in better reading skill.

d. The *p*-value tells the probability of getting the observed results, if the new program does not result in better reading skill.

This item is classified in the **Analysis of the Sample** category and assesses student understanding of the role of *p*-values in the ***logic and reasoning*** of statistical hypothesis testing. Answer choice *d* is the correct answer. Answer choice *c* is similar to option *d* in that it attaches the *p*-value to the sample. However it is different from answer choice *d* in that it states that the probability is conditioned on the null hypothesis being false. Students might choose this option if: (a) they thought the probability was conditioned on the null being true; and/or (b) they thought the probability was conditioned on whichever hypothesis supports the data (note that the data support the fact that there is a difference in reading programs – this might have led a student to choose this option). Answer choices *a* and *b* do not attach *p*-value to the sample. These options test whether a student understands *p*-value to be: (1) the probability that the null is/is not true; and/or (2) the probability that the observed difference does/does not describe the true nature of things.

Only 11.5% of the participants chose the correct answer while 45.2% of the participants chose *c*, and 29.8% of the participants chose *a*. These results suggest that students do not have strong understandings of *p*-value: what it represents and how it is used.

In summary, the analysis of the results in the **Analysis of the Sample** and **Decision Rule** categories indicates that students do not have strong understandings of the statistical concepts of sampling distributions, significance, and *p*-value. These are all ideas and concepts that must be referenced when performing a statistical hypothesis test. The relatively high scores earned on the course exam, however, indicate that students were able to use the statistical tables (which give probabilities associated with sampling distributions), the level of significance, and *p*-values to solve the problems found there.

On the other hand, analysis of student response on the multiple-choice assessment indicates that students do not understand the underlying concepts. This is an example where the relationship between students' understandings of the procedures and the concepts and logic is not strong. We next turn to the analysis of results from the **Conclusion** and **Recognizing Applicability** items to gain further insight into student understanding of the logic and reasoning of statistical hypothesis testing.

Conclusion and Recognizing Applicability Categories

The **Conclusion** and (in part) the **Recognizing Applicability** categories assess students' understanding of the conclusions that can be made as a result of statistical hypothesis testing and how those conclusions are of value. Statistical hypothesis testing is a method by which an inference about a population can be drawn from analysis of a sample. With that inference comes a degree of uncertainty. Statistical hypothesis testing provides a means of quantifying that uncertainty and, because this is the case, it is a powerful method of inference. Performance on items within these categories indicates that students have misunderstandings about these issues. Given sample information and/or results from a statistical hypothesis test, introductory students struggle to draw appropriate inferences about the population. In particular, they believe that statistical hypothesis tests establish the "truth" of a hypothesis and that sample information directly translates to population characteristics.

Hypothesis Testing and the "Truth" of a Hypothesis. Statistical hypothesis testing relies on indirect reasoning to determine whether or not to reject the null hypothesis. Regardless of the decision to reject or fail to reject, statistical hypothesis testing does not provide a proof of either the null or the alternative hypotheses. Nor does it establish the probability that either the null or the alternative hypotheses is true. It does, however provide a means of quantifying the uncertainty associated with a decision to reject or fail to reject the null hypothesis. Using the level of significance, one is able to make a claim about the probability that a decision to reject the null hypothesis is incorrect if the null hypothesis is, indeed, correct. In addition, if one fails to reject the null hypothesis it is possible to calculate the probability that this decision is in error if, indeed, the null hypothesis is not true. This calculation relies on a specified effect size and the parameter value under the alternative hypothesis. This understanding is essential to statistical hypothesis testing as it provides a foundation for sample analysis and for the kinds of conclusions that may be made about that sample.

Item number 3 is classified in the **Recognizing Applicability** category and assesses whether students understand statistical hypothesis testing as a means for making a decision about a population based on information from a sample and that the uncertainty associated with that decision can be quantified.

Figure 4.9

Item Number 3, Multiple-Choice Assessment

3. Which of the following statements is the ***best justification*** for using a statistical hypothesis test to answer the question: *Are female students as successful as male students in college mathematics*?

a. It allows you to talk about uncertainty associated with your decision.

b. It allows you to use only a sample of students to prove something for all students.

c. It allows you to calculate means and use statistical tables to answer the question.

d. It allows you to find and prove the answer using mathematical calculation.

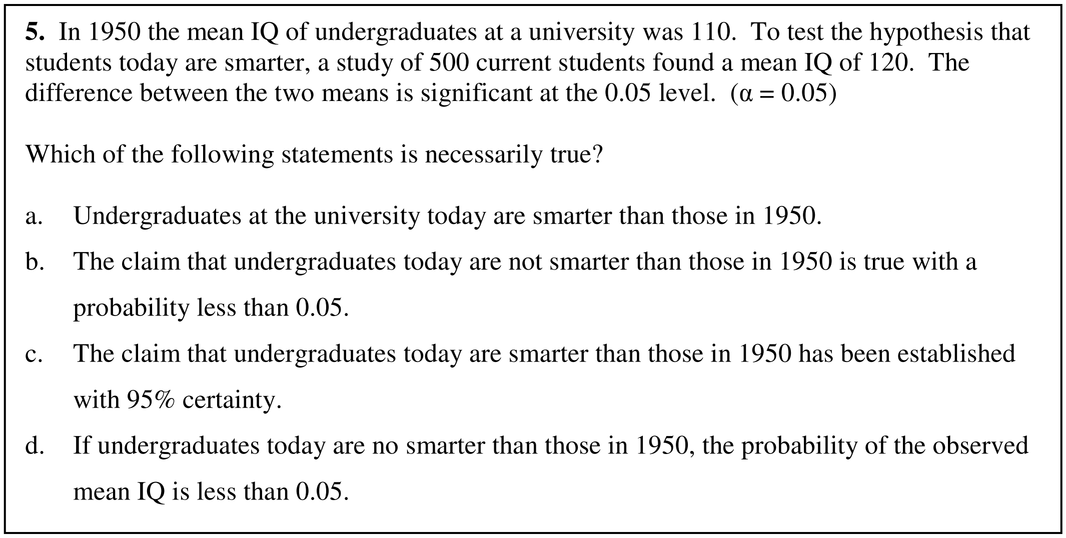
Answer choice *a* is correct and answer choices *b* and *d* were written to test whether participants believe statistical hypothesis testing to be a proof. Answer choice *c*, while true, is not the *best* justification. If a student chooses this option over choice *a*, *b*, or *d* this is an indication, that the student understands that statistical hypothesis testing is not a proof. Additionally, he/she either doesn't understand statistical hypothesis testing as a measure of uncertainty associated with a decision or he/she thinks that value in doing statistical hypothesis testing is in using means and statistical tables, rather than providing a way to talk about uncertainty. The latter could, potentially, result from a very procedural understanding of statistical hypothesis testing.

Only 11.5% of the participants chose *a*, 35.6% of the participants chose *b*, 28.8% of the participants chose *d*, and 24% of the participants chose answer choice *c*. Over half of the participants chose one of *b* or *d*. This result provides evidence that students understand statistical hypothesis testing to be a proof, rather than a process by which to make a claim and to talk about the uncertainty associated with making that claim. When given the choice from among various options, they do not associate statistical hypothesis testing with providing a means to make a decision about a claim for which there is always uncertainty, while quantifying the potential error associated with that decision if, in reality, the assumed claim is/is not true. Introductory statistics students do not value statistical hypothesis testing as an inferential method used to "make sense" of variable data.

The results from item number 5 provide further insight into this issue. Classified in the **Conclusion** category, this item assesses whether students understand that statistical hypothesis testing gives information about the degree that the sample provides a counterexample to an assumed (null) hypothesis. It assesses whether students understand that the results of statistical hypothesis tests do not provide proofs of claims, nor do they provide the probability that a particular claims are true or untrue.

Figure 4.10

Item Number 5, Multiple-Choice Assessment



5. In 1950 the mean IQ of undergraduates at a university was 110. To test the hypothesis that students today are smarter, a study of 500 current students found a mean IQ of 120. The difference between the two means is significant at the 0.05 level. ($\alpha = 0.05$)

Which of the following statements is necessarily true?

a. Undergraduates at the university today are smarter than those in 1950.

b. The claim that undergraduates today are not smarter than those in 1950 is true with a probability less than 0.05.

c. The claim that undergraduates today are smarter than those in 1950 has been established with 95% certainty.

d. If undergraduates today are no smarter than those in 1950, the probability of the observed mean IQ is less than 0.05.

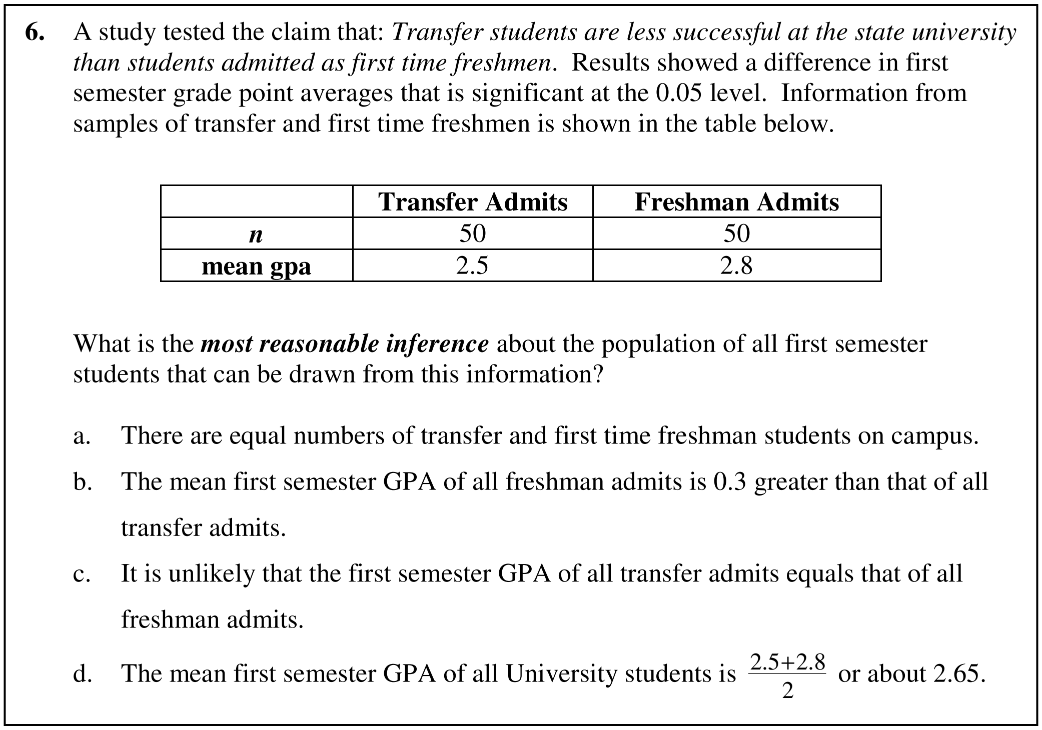
Although answer choice *d* is correct, only 6.7% of the students chose it. An overwhelming number of participants chose answer choice *c*: 87.5% of the participants. Very few chose *b* or *a*. These results indicate that students believe the level of significance to be associated with the probability that the null (or alternative) hypothesis is true.

Combined, the results reported for these two items provide evidence that introductory statistics students associate statistical hypothesis testing with a means to establish the "truth" of a given hypothesis and/or as a means to determine the degree to which a hypothesis is true. They do not understand statistical hypothesis testing to rely on indirect reasoning to test the feasibility of an assumed null hypothesis. And, they do not understand that the conclusions to statistical hypothesis tests have some degree of uncertainty associated with them.

Sample Statistics and Population Parameters. In addition to difficulty interpreting a conclusion, introductory statistics students seem to struggle with the idea of inference, in general. In statistical hypothesis testing, sample statistics are used to make an inference about a claim about a population parameter. Though sample statistics are used to make a claim about the feasibility of a hypothesis about population parameters, they are not direct measures of those populations' parameters. The inference from a sample to a population is not a direct conversion of sample statistic to population parameter. It is an inference about the relative magnitude of the population parameter. Unfortunately, students do not have this understanding, as evidenced by the results of item number 6.

Figure 4.11

Item Number 6, Multiple-Choice Assessment



6. A study tested the claim that: *Transfer students are less successful at the state university than students admitted as first time freshmen*. Results showed a difference in first semester grade point averages that is significant at the 0.05 level. Information from samples of transfer and first time freshmen is shown in the table below.

	Transfer Admits	**Freshman Admits**
n	50	50
mean gpa	2.5	2.8

What is the ***most reasonable inference*** about the population of all first semester students that can be drawn from this information?

a. There are equal numbers of transfer and first time freshman students on campus.

b. The mean first semester GPA of all freshman admits is 0.3 greater than that of all transfer admits.

c. It is unlikely that the first semester GPA of all transfer admits equals that of all freshman admits.

d. The mean first semester GPA of all University students is $\frac{2.5+2.8}{2}$ or about 2.65.

Answer choice *c* is the correct option and 36.5% of the participants chose this option. The other answer choices were written to assess whether students believe that there is a direct translation from the sample statistics to the population parameters. Only a few participants chose option *a*. Answer choices *b* and *d*, however, were fairly popular with 34.6% of the participants choosing *b* and 21.2% of the participants choosing *d*. These results indicate that students believe that a direct translation from sample to population is a valid inference to make. This belief is not in alignment with the logic and reasoning of statistical hypothesis testing. Samples vary, as do their summary statistics. Therefore, it is not correct to infer that sample statistics directly translate to population parameters.

The results reported in this section are interesting especially when compared with the results of the course exam. On the course exam and in well-defined statistical hypothesis testing problems, students are asked to make a "concluding statement." They must state whether or not the null hypothesis should be rejected. Relatively high scores on the course exam indicate that students are able to make the correct concluding statement. However, as the preceding analysis indicates, students do not necessarily understand the logic and reasoning that supports that concluding statement, nor do they understand what that concluding statement really "means," so to speak. The concept of inference is problematic and students do not consistently make valid inferences about sample information.

Research Sub-question Number 3

Analysis of the results associated with the items classified as ***method and context*** is useful to better describe how introductory students understand the relationship between the method of statistical hypothesis testing and the context within which it is employed. Table 4.8 gives aggregate data for items in this category.

Table 4.8

Method and Context Category, Multiple-Choice Results

Item	Framework Category	Percent Correct	Average
1	Recognizing Applicability	43.27	40.00
2	Recognizing Applicability	61.54	
10	Collect a Sample	45.19	
12	Implication for Practice	36.54	
14	Implication for Practice	13.46	

On average, 40% of these items were answered correctly. There is evidence, therefore, that introductory statistics students have a stronger understanding of the relationship between statistical hypothesis testing and the context in which it is employed than they do of the overall logic and reasoning associated with the method.

The items labeled ***method and context*** are classified in the **Recognizing Applicability**, **Collect a Sample**, and **Implication for Practice** categories of the Framework for Assessing Understanding. Though not a strong performance by common standards, performance on the items classified in the **Recognizing Applicability** and **Collect a Sample** categories (item numbers 1, 2, and 10) of the *method and context* grouping was relatively high while performance on the **Implication for Practice** items (numbers 12 and 14) was relatively low. Analysis of the frequencies associated with the various answer choices for these items gives some insight into student understanding of these pieces of statistical hypothesis testing.

The strongest performance was on item number 2 (of the **Recognizing Applicability** category). This item assesses whether students understand that statistical hypothesis testing may only be used to address research questions that can be answered through analysis of a *measure*. That is, it must be possible to quantify the question. Given the results, it seems that students understand this to be the case. Over half of the participants (61.5%) choose the correct answer, option *d*.

Figure 4.12

Item Number 2, Multiple-Choice Assessment

2. Which of the following actions is the most ***important first step*** in designing a statistical hypothesis test to answer the question: *Are out-of-state students more successful at the state university than students who are in-state residents*?

a. Agree on a statistical test to compare the groups.

b. Agree on a sample size from each population.

c. Identify available statistical software.

d. Agree on a way of measuring student success.

After option choice *d*, the distractor most chosen was answer choice *b*, with 25% of the participants choosing that option. Though *d* is the most important thing to determine first, option *b* is important. However, the test can not even be conducted, if the researcher cannot find a way to quantify his/her question.

Fairly strong performances were reported on items 1 (of the **Recognizing Applicability** category) and item 10 (of the **Collect a Sample** category). Though performance on these items was relatively high (a little under half of the participants answered them correctly), it is interesting to note that, for each item there was a single attractive distractor. That is, the majority of participants either chose the correct option or they chose only one of the distractors.

Item number 1 assessed student ability to identify situations where hypothesis testing might be used to answer questions of interest: The question is answerable by a measure of the population, it is only feasible to test a sample of the population, and there are two contradictory hypotheses that can answer the question.

Figure 4.13

Item Number 1, Multiple-Choice Assessment

1. Which of the following questions is ***most likely*** to be answered by a study that requires statistical hypothesis testing?

a. Do athletes have a lower GPA than other students?

b. What equation predicts a student's freshman GPA from his/her SAT score?

c. What are typical costs for full-time resident students in U. S. colleges?

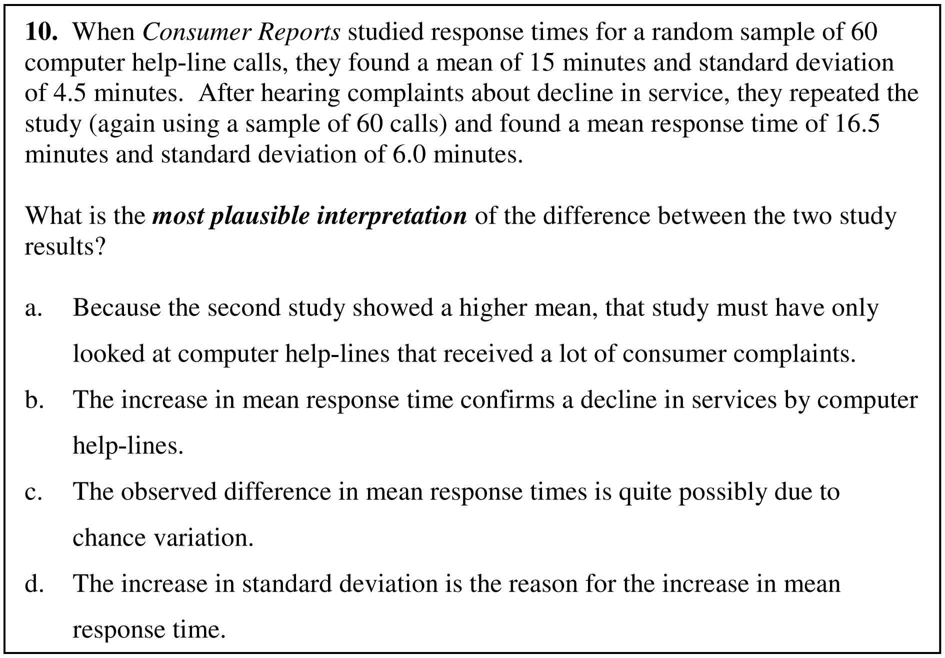
d. Do the 12:00 noon sections of STAT 100 perform better than the 2:00 p.m. sections this semester?

Performance on this item was relatively high with 43.3% of the participants choosing the correct answer. However, over half of the participants chose answer choice *d*. This result indicates that, while introductory statistics students understand statistical hypothesis testing to be useful in comparing two groups, they do not necessarily understand the power of the test in making *inferences* from a sample to a population.

Item number 10 was written to assess whether students understand that samples vary and, therefore, in order to make a claim about a population, the sample must be randomly chosen and must be representative of the population.

Figure 4.14

Item Number 10, Multiple-Choice Assessment



10. When *Consumer Reports* studied response times for a random sample of 60 computer help-line calls, they found a mean of 15 minutes and standard deviation of 4.5 minutes. After hearing complaints about decline in service, they repeated the study (again using a sample of 60 calls) and found a mean response time of 16.5 minutes and standard deviation of 6.0 minutes.

What is the ***most plausible interpretation*** of the difference between the two study results?

a. Because the second study showed a higher mean, that study must have only looked at computer help-lines that received a lot of consumer complaints.

b. The increase in mean response time confirms a decline in services by computer help-lines.

c. The observed difference in mean response times is quite possibly due to chance variation.

d. The increase in standard deviation is the reason for the increase in mean response time.

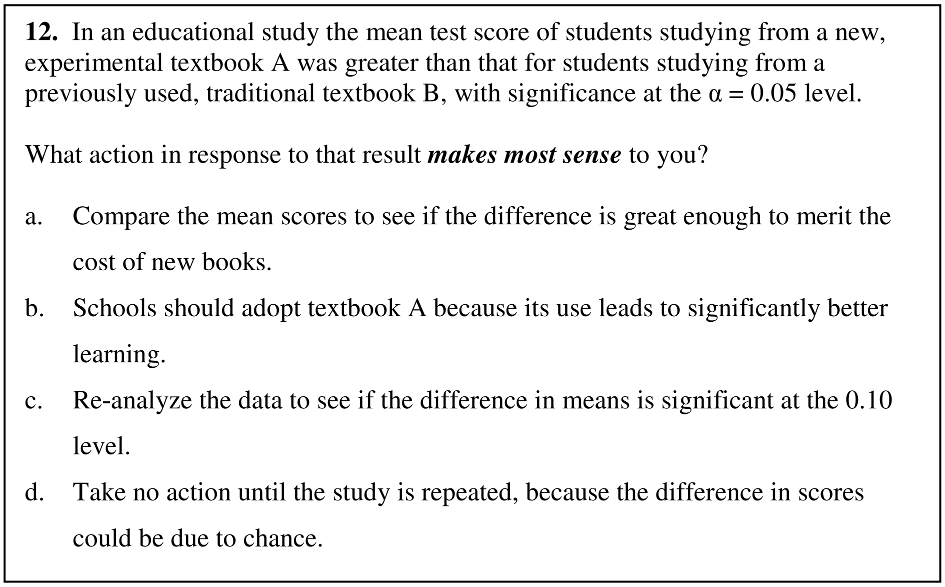
Answer choice *c* is the correct answer and 45.2% of the participants chose this option. However, many of the participants (32.7%) chose answer choice *b*. This result provides evidence that many introductory statistics students do not understand the variability associated with samples, and believe that only one sample is needed to confirm a hypothesis. If this were the case, then statistical hypothesis testing would not be necessary, as analysis of summary statistics of a sample would be all that is needed to confirm a claim.

Weaker performances were associated with items 12 and 14 from the **Implication for Practice** category. These items assess whether students understand that statistical significance does not necessarily imply practical significance. In addition to statistical significance and *p*-value, policy makers must consider a variety of factors when making a decision. Sample size, effect size, and Type I and Type II errors are all important factors to consider when making potentially costly decisions.

Item number 12 assesses student understanding of these issues and, for this item, participants chose a variety of answer choices.

Figure 4.15

Item Number 12, Multiple-Choice Assessment



12. In an educational study the mean test score of students studying from a new, experimental textbook A was greater than that for students studying from a previously used, traditional textbook B, with significance at the $\alpha = 0.05$ level.

What action in response to that result ***makes most sense*** to you?

a. Compare the mean scores to see if the difference is great enough to merit the cost of new books.

b. Schools should adopt textbook A because its use leads to significantly better learning.

c. Re-analyze the data to see if the difference in means is significant at the 0.10 level.

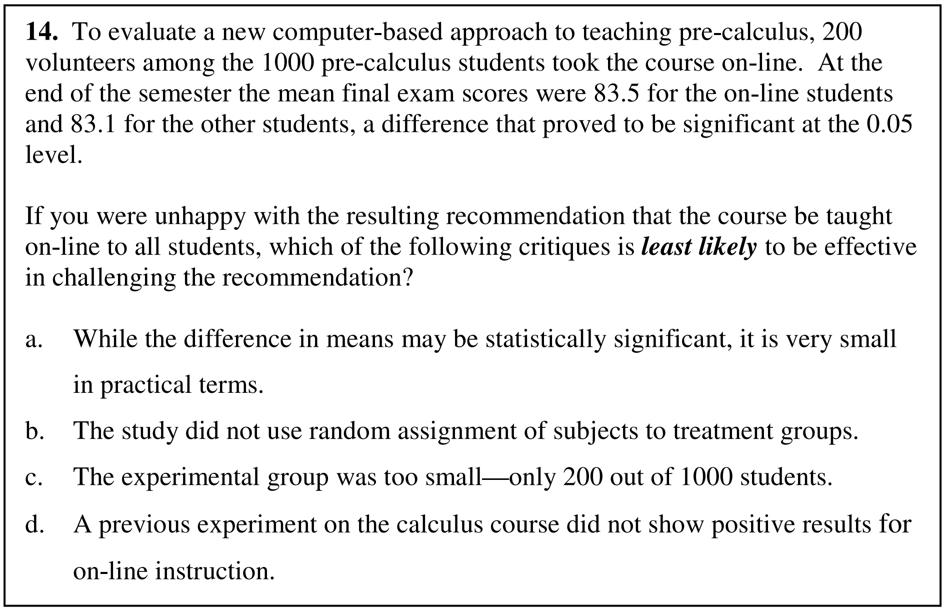
d. Take no action until the study is repeated, because the difference in scores could be due to chance.

Answer choice *a* is correct and 36.5% (or 38) of the participants chose this option. Answer choice *b* assesses whether students believe that, when statistical significance has been achieved, there is no need to consider other factors before action. Answer choice *c* relates to item number 11 in assessing student understanding of significance level. Answer choice *d* assesses whether students understand that statistical hypothesis tests provide a powerful means of drawing an inference about large populations from which it is too costly and time consuming to collect information on all members. It is also costly and time consuming to collect *sample* information from such populations. This is the case for the study described in item number 12. Given that the sample size was sufficient (and there is no reason to believe it is not) the study should not be repeated as data collection would be costly. Thus, answer choice *a* is the best option. However, answer choices *b*, *c*, and *d* were very popular with 20.2% of the participants choosing answer choice *b*, 19.2% of the participants choosing answer choice *c*, and 24% of the participants choosing answer choice *d*. The variability in responses indicates that students do not understand how to use the results of a statistical hypothesis test in context.

This claim is confirmed by the results of item number 14, where students were asked to consider issues that may/may not be effective in challenging the results of a study for which statistical significance was achieved.

Figure 4.16

Item Number 14, Multiple-Choice Assessment



14. To evaluate a new computer-based approach to teaching pre-calculus, 200 volunteers among the 1000 pre-calculus students took the course on-line. At the end of the semester the mean final exam scores were 83.5 for the on-line students and 83.1 for the other students, a difference that proved to be significant at the 0.05 level.

If you were unhappy with the resulting recommendation that the course be taught on-line to all students, which of the following critiques is ***least likely*** to be effective in challenging the recommendation?

a. While the difference in means may be statistically significant, it is very small in practical terms.
b. The study did not use random assignment of subjects to treatment groups.
c. The experimental group was too small—only 200 out of 1000 students.
d. A previous experiment on the calculus course did not show positive results for on-line instruction.

Here, the correct answer is choice *d*. The other options present legitimate challenges to the results of the study. Only 14 (or 13.5%) of the participants chose the correct answer. Answer choices *a*, *b*, and *c* were all popular with 34 (or 32.7%), 26 (or 25%), and 29 (or 27.9%) of the participants choosing each, respectively. These results are questionable, however, as the students may have inadvertently interpreted the question to ask for the "most likely" rather than the "least likely".

Overall the results in this category indicate that, though introductory statistics students have a better grasp of the role of context in statistical hypothesis testing than they do of the reasoning and logic, this understanding is not strong. It does seem that students understand, to some degree, that various factors within the context must be considered when (1) setting up and conducting a statistical hypothesis test and when (2) using the results to influence action.

Summary and Conclusions: Quantitative Phase

The analysis of the quantitative results presented in this chapter provides evidence that, though students in traditional, introductory statistics courses do well on exams constructed with traditional, well-defined statistical hypothesis testing problems, they do not necessarily have the desired understandings of the concepts, logic, and uses of the method. From the stranded perspective of proficiency offered by *Adding It Up* (Kilpatrick et. al, 2001) these results indicate that a high degree of *procedural fluency* does not necessarily mean high degrees of *conceptual understanding*, *strategic competence*, *adaptive reasoning*, and *productive disposition* (at least, in regard to having an understanding of the value of the method) with respect to statistical hypothesis testing.

In particular, the results indicate that introductory statistics students do not understand the role of probability and inference in statistical hypothesis testing. This difficulty is most likely enhanced by demonstrated weak understandings of sampling distributions and of the role that sampling distributions play in statistical hypothesis testing. They do not seem to understand the degree to which samples vary and believe that statistical hypothesis tests prove a given hypothesis and/or provide a measure of the degree to which a given hypothesis is true. In addition, introductory statistics students believe that sample statistics provide direct measures of the population. They do not understand that probability is used in the inferential process to quantify the uncertainty associated with the stated conclusion and do not value statistical hypothesis testing as an *inferential* method used to study large populations for which only sample information can be obtained. Though they can correctly state the hypotheses and understand the method to be useful in comparing two populations, introductory statistics students do not understand that the power of the method lies in its ability to draw inferences about large populations through analysis of only a sample. Finally, the data provide evidence that the misunderstandings students have concerning the role of probability and inference influence the ways in which they interpret and apply the results of a statistical hypothesis test in “real world” contexts.

Though the items and distractors were carefully constructed and the analysis seems reasonable, the conclusions here are limited. As is the case with all multiple-choice assessments, inferences about student thinking based on information obtained from these instruments may be incomplete or even flawed. Therefore, it is important to interview students who represent a range of performance patterns on the two quantitative assessments to gain more insight into their thinking. The results and analysis of this qualitative phase of the study are presented in the next chapter.

CHAPTER 5

QUALITATIVE RESULTS AND ANALYSIS

In this chapter, results of the qualitative phase of the study will be presented and analyzed. Eleven students representing the various performance patterns in the quantitative phase participated in follow-up interviews. Analysis of student responses yielded information that addresses the three identified research sub-questions and provides insight into introductory statistics students' understandings of statistical hypothesis testing. The chapter is organized into three sections: (1) outline of the interview design, (2) presentation of the data and the analysis of that data, and (3) summary of the data and conclusions associated with the qualitative phase of the study.

Interview Design

The follow-up interview was designed to provide more insight into student thinking about statistical hypothesis testing than could be provided on the multiple-choice assessment and to provide a means of data triangulation to validate the claims made in the quantitative phase of the study. Interviewees were asked to explain their thinking on various items from the multiple-choice assessment, why they chose the option they did as well as why they did not chose the other options. Additional questions were asked to probe each individual's thinking about each item. Interviewees were provided with pencil and paper and could use these tools as necessary to explain their thinking. They were also told that they could change their answers at any point during the interview.

Every attempt was made to insure that the interview was conducted as a natural conversation between interviewer and interviewee.

From the pilot study, it was determined that the follow-up interview should last no longer than an hour. Introductory statistics students begin to fatigue after having explained their thinking for longer than an hour. From the pilot study, it was also determined that asking students to explain their thinking for 9 multiple-choice items was a reasonable request for an hour-long, follow-up interview. Thus, 5 of the 14 multiple-choice items were eliminated from the interview protocol.

Because the follow-up interview was designed to extend the analysis conducted in the quantitative phase, the results of the assessments were used to inform the design of the follow-up interviews. The 9 items included in the follow-up interview were strategically chosen to be representative of the categories of the Framework for Assessing Understanding of Statistical Hypothesis Testing. They were also chosen to maximize the potential for variability that might exist among student responses (i.e. choosing items for which performance was low). The results of the assessment, therefore, were useful in helping to determine which 9 items should be included. Additionally, as it was not possible to interview all 104 participants from the quantitative phase, the results of the assessments were used to inform the choice of interviewees. Interviewees were chosen so that they were representative of the groups of students participating in the quantitative phase. This selection process allowed for the potential for variability among responses to be maximized.

Item Elimination

In order to address the second and third research sub-questions, more than half of the ***logic and reasoning*** items (6 of 9 items) and more than half of the ***method and context*** items (3 of 5 items) were included in the follow-up interview. Furthermore, these items were chosen to be representative of each of the 7 categories identified by the Framework. Finally, these items were chose to maximize the potential for variability among responses. Items for which performance was low best served that purpose.

Within the ***logic and reasoning*** category, item numbers 7, 8, and 13 were eliminated. Item number 7 is classified in the **Generating Statistical Hypotheses** (**GH**) category and, of the ***logic and reasoning*** items, it showed strongest performance. Because performance on this item was strong and because another item representing the **GH** category could be included, item number 7 was eliminated. Item numbers 8, 9, and 13 are all classified in the **Analysis of the Sample** category. Item numbers 8 and 9 both address the concept of sampling distributions and item number 13 addresses the concept of *p*-value. Responses were more varied for item number 9 than they were for item numbers 8 and 13. Additionally, it was possible within the context of item number 9 to ask probing questions that addressed the ideas associated with item numbers 8 and 13. Therefore, item numbers 8 and 13 were eliminated. Ultimately, item numbers 3, 4, 5, 6, 9, and 11 from the ***logic and reasoning*** category were included in the follow-up interview.

Within the ***method and context*** category, item numbers 2 and 14 were eliminated. Item number two is classified in the **Recognizing Applicability** (**RA**) category and, of the ***method and context*** items, it showed the strongest performance. Because performance on this item was strong and because two other items representing the **RA** category could be included, it was eliminated. Additionally, item number 14 (classified in the **Implication for Practice** category) was eliminated in favor of item number 12 from the same category. Performance on item number 14 was low and may have been due to misinterpretation of the item (see Chapter 4). It was, therefore, expected that more insight about student thinking about **Implication for Practice** would be gained by including item number 12, rather than 14. Ultimately, item numbers 1, 10, and 12 from the ***method and context*** category were included in the follow-up interview.

As a result of this elimination process, nine items were included in the follow-up interview. These items (1) are representative of the various categories outlined in the Framework, (2) maximize the potential for variability among student responses, and (3) address research sub-questions 2 and 3. The items and their classifications are shown in Table 5.1.

Table 5.1

Items Used in the Interview and Their Classifications

Item Number	Framework Category	Research Sub-question Category
1	Recognizing Applicability	Method and Context
3	Recognizing Applicability	Logic and Reasoning
4	Generating Statistical Hypotheses	Logic and Reasoning
5	Conclusion	Logic and Reasoning
6	Conclusion	Logic and Reasoning
9	Analysis of the Sample	Logic and Reasoning
10	Collect a Sample	Method and Context
11	Decision Rule	Logic and Reasoning
12	Implication for Practice	Method and Context

Given this set of 9 items, quantitative results for each item were analyzed and used to determine whether there were particular issues and/or questions that should be raised in the follow-up interview. A list of these additional questions and issues can be found in appendix C.

Participants

In order to maximize the variability among participants, interviewees were chosen to represent the various performance patterns on the course exam and multiple-choice assessment. As described in Chapter 3, these groups were constructed using the quartiles into which scores on the course exam and multiple-choice assessment fell. Three groups of students were of particular interest for the follow-up interview: those who scored in the top quartile of scores on both assessments (**HH**); those who scored in the top quartile of scores on the course exam, but in the bottom quartile of scores on the multiple-choice assessment (**HL**); and those who scored in the bottom quartile of scores on the course exam, but in the top quartile of scores on the multiple-choice assessment (**LH**). The groupings are shown in Table 5.2.

Table 5.2

Groupings for Interviews

Category	Course Exam Score (%)	Multiple-Choice Score (%)	Number of Participants Quantitative Phase	Number of Participants Qualitative Phase
HH	Greater than or equal to 90	Greater than or equal to 43	11	3
HL	Greater than or equal to 90	Less than or equal to 21	7	4
LH	Less than or equal to 58	Greater than or equal to 43	5	4

Once students were placed into appropriate groups, interviewees were chosen at random from each group. If these students consented to the follow-up interview, they were invited to participate. If a potential interviewee declined, another was chosen at random until the desired number of participants was obtained. Unfortunately, in the case of the **HH** and **LH** groups, it was not possible to schedule interview times with the required number of individuals. In both cases, however, an effort was made to locate individuals who earned scores that were close to the desired boundaries. As a result, one individual who obtained an 86% on the course exam and a 71% (the highest score earned, overall) on the multiple-choice assessment was included in the **HH** group. And, an individual who earned a 65% on the course exam and a 57% on the multiple-choice assessment (the second highest score earned, overall) was included in the **LH** group.

There was diversity among the group of interviewees with respect to gender, race, major, and year in school. Table 5.3 illustrates this diversity. Note that interviewees are identified by their survey number.

Table 5.3

Participant Characteristics

Interviewee Number	Group Classification	Gender	Race	Major	Year in School
15	HH	Female	White	Conservation Science	Junior
112	HH	Male	White	Journalism	Freshman
132	HH	Female	White	Pre-nursing	Junior
77	HL	Female	Asian	Pre-nursing	Senior
122	HL	Female	Asian	Family Sciences	Junior
169	HL	Male	Asian	Journalism	Freshman
172	HL	Female	White	Journalism	Freshman
29	LH	Female	Asian	Pharmacy	Junior
81	LH	Male	White	Kinesiology	Sophomore
191	LH	Male	White	Kinesiology	Sophomore
192	LH	Female	Black	Pre-nursing	Junior

As much as possible, the follow-up interviews were scheduled before students left campus for winter break. Due to scheduling conflicts, however, three of the eleven interviews were conducted during the winter session, in January.

Implementation

To situate the interview, and to gain insight into the interviewees' general impressions of statistical hypothesis testing, each interview began by asking the interviewee to describe statistical hypothesis testing and to explain how it could be used. Then, over the course of an hour, each interviewee was given the opportunity to explain his or her reasoning about each of the 9 identified multiple-choice items. Every attempt was made to ask similar probing questions with respect to those items. However, in some cases, as the hour drew to an end, there was not time for additional probing questions.

Each interview was audio-taped and transcribed. Within groups (**HH**, **LH**, and **HL**), the data was analyzed for commonality as well as uniqueness of answers. Analysis then continued across the groups in search of common themes that transcended performance patterns. Though it was hypothesized that different themes would arise for each group, with only some commonality between groups, this was not the case. The groups were not dissimilar from each other in their reasoning about the items in any recognizable way. This result, however, is not entirely unreasonable given that performance on the multiple-choice assessment was low for each of the groups. Because overall performance was low on the multiple-choice assessment, the groups appeared very similar when asked about their reasoning for the items on that instrument. Therefore, there will be no distinction between the groups of interviewees in the presentation of the data and analysis in this chapter.

Overall, the interviews were useful in providing insight into student thinking and in triangulation of data sources. The results and analysis of the interviews will be presented in the next section.

Data and Analysis

Data collected in the qualitative phase gives information about introductory statistics students' understandings of statistical hypothesis testing. Specifically, this data addresses the following research sub-questions:

1. *What is the relationship between introductory students' understanding of the procedures and the concepts, logic, and uses of statistical hypothesis testing?*
2. *What are the understandings that introductory students have of the overall logic and reasoning of statistical hypothesis testing?*
3. *What are introductory students' understandings of the relationship between the method of statistical hypothesis testing and the context in which it is employed?*

and provides information about the overriding research question that guided data collection in this study: What are the understandings of statistical hypothesis testing held by students who have completed a traditional, introductory course in statistics at a large university? The data collected in the follow-up interviews confirmed that which was found in the quantitative phase of the study. Introductory statistics students do not have strong, connected understandings of statistical hypothesis testing. Additionally, this data points to particular aspects of the method and its uses that are problematic for introductory statistics students and illustrates the understandings that these students *do* have of statistical hypothesis testing. In this section, the data collected in the follow-up interviews and analysis of that data will be presented.

In order to situate the presentation of the analysis, student responses to the lead-off question from the interview will be presented first. Not only does this lead-off question situate the analysis, but it is also useful to address the guiding research question of the study. Then, the results and analysis of the data with respect to the research sub-questions will be presented.

Research sub-question number 1 is an inquiry into the relationship between student understanding of the procedures and the concepts, logic, and uses of statistical hypothesis testing. In explaining their reasoning for the answers they chose in each of the ***logic and reasoning*** and ***method and context*** items, interviewees often referred to their knowledge of procedures as well as to their knowledge of the concepts, logic, and uses. Thus, for this qualitative analysis, data used to address research sub-question number 1 is *subsumed within* data used to address research sub-questions 2 and 3. For this reason, the analysis of data with respect to research sub-questions 2 and 3 will be presented before the analyses with respect to research sub-question number 1.

In each section, a summary of student responses will be provided and followed by an analysis of those responses. Excerpts from individual interviews are included. In these excerpts, it should be noted that statements preceded by an "M" were made by the interviewer and statements preceded by an "I" were made by the interviewee.

Overall, analyses of the three research sub-questions across the groups of participants provides insight into the way in which introductory statistics students understand statistical hypothesis testing and extends the analysis that was conducted in the quantitative phase of the study.

Overall Description and Impression of Statistical Hypothesis Testing

At the beginning of each follow-up interview, interviewees were asked to describe hypothesis testing and how it might be used. Analysis of student responses to this question indicates that introductory statistics students have somewhat different impressions of statistical hypothesis testing. For many of the interviewees, these differences were similar to those outlined by approaches taken by Fisher and by Neyman and Pearson. For others, their description of statistical hypothesis testing was very different from either Fisher or Neyman and Pearson. The way in which students describe statistical hypothesis testing provides insight into their understanding of the method and its uses.

Interviewee Descriptions: A Summary

When asked to describe statistical hypothesis testing and how it might be used, two interviewees described it as a process by which one attempts to find evidence against an assumed null hypothesis. For these interviewees, the null hypothesis represents the status quo and the alternative hypothesis represents that which the researcher is trying to prove. Thus, hypothesis testing allows researchers to test whether the null (status quo) is, indeed, true and if not, then confidence is placed in the alternative hypothesis. This understanding of statistical hypothesis testing is reflected in the following quotes:

> **Excerpt: Interview 81**
> I: Yeah, it's kind of you're testing to see if…if the alternative's actually what's correct, I guess. You're testing it…it's not like you're testing it against each other. It's more like you're just…you're…you're, you already knew the null hypothesis to be right. You're trying to see if it's still right…or if it's sort of changed or something like that.
>
> **Excerpt: Interview 191**
> I: Um, I remember it's establishing, ah, the success failure, so a null hypothesis and an alternative hypothesis. Um, I'm trying to think.

discussion continues

M: Success and failure, you mean [I: Yeah] like…can you say more about that, what you mean by that?

I: Um, well, there's, there's two tests, one, ah, that it's based on a question, so, um, you'll have… you'll have one hypothesis that states one thing, like ah, I remember we did some…usually it would be *u* is equal to a certain value, um, and the alternative hypothesis would be, ah, *u* is less than, or greater than, or not equal to certain value, um, depending on the question, so if the question asked, um, if something was less than, ah, a certain value, you'd use that for your, ah, alternative hypothesis. [M: Ok]. Um, yeah, it was, it was based on a question. Ah, your a, alternative hypothesis was…your null hypothesis I think was usually only like a standard, if it's equal to something.

M: Ok. Ok, and so you have these two things and then the success failure comes in –

I: Based, based on whether…whether you prove if the, ah, null hypothesis is, is correct. [M: Ok] If it's correct, then it's success. If it's false, then it's a failure.

An understanding of hypothesis testing as a test of the null hypothesis aligns with the logic Fisher used in his thinking of statistical hypothesis testing. Though Interviewees 81 and 191 understand statistical hypothesis testing as a test of the null hypothesis, it is not clear that these individuals understand statistical hypothesis testing as an application of (modified) proof by contradiction. However, one interviewee, number 112, stated this logic very clearly.

Excerpt: Interview 112
I: Pretty much you compare…like…you have two different ideas that there's H0 and H1 and H0's like what the already accepted thing is and H1's what you're trying to prove. But, in order to prove H1 you have to be able to disprove H0. And, there's different alpha values you can use, but usually a good one to use is like 95. Ninety-five is like alpha value 0.05, 0.25.

His comment indicates that, on some level, he understands that statistical hypothesis testing relies on indirect reasoning to reach a conclusion.

Other interviewees (5 of the 11 interviewees) expressed an understanding of statistical hypothesis testing as a test in which one hypothesis "competes" against another. The following quotes reflect this understanding.

Excerpt: Interview 77

I: Um…I guess pretty much testing…you're trying to test the claim and whether it's true or not….it's like the null hypothesis and the alternative hypothesis. And, so you want to find out, like, which one to, I guess, which hypothesis is true.

Excerpt: Interview 172

I: Sure, um, well…the way I understand it is that hypothesis testing is used to um…I guess it…decide whether…ah…one measurement, I guess, is correct or not and, ah….I don't know. The way most of the problems that we've been presented with work is…ah…they'll have sort of a known, ah, status quo, I guess of a…so to speak…that, ah…this many students got, failed this class or this many students passed or something…and then they'll be a new study or a new thing that…says that's a different type of number and you test to see which one is correct under a certain…ah…significance level. And, I know how it can change based on the significance level that you use.

This approach to hypothesis testing aligns with that described by Neyman and Pearson. In this approach, the null and alternative hypotheses are tested against each other to determine which hypothesis the sample supports. It is not clear, however, that the interviewees understand the way in which these two hypotheses are tested against each other. Based on their statements, we do not know if the interviewees understand how the data collected provides evidence for or against the null or alternative hypotheses.

Other interviewees described hypothesis testing in slightly different terms. For example interviewee number 15 described it as a method researchers use to show "that they just aren't making it up. They have all this data that proving like that there actually is a conclusion that they can come across and that they're not just kind of like pulling it out of thin air and throwing it out there" (Interview 15, 2008). Interviewee 15 understands hypothesis testing to be means by which researchers can prove their claim to a larger community. Her statement, however, does not indicate whether or not she understands how this method "proves" a claim to a given community of individuals.

Interviewee number 122 described it as follows:

Excerpt: Interview 122

> I: Well, it usually to make, uh, kind of get, ana – like, kind of analysis of data, like, they're just for random samples and they're raw data. So you try and use that to formulate a hypothesis regarding to probably different population from another or a certain topic that you're trying to figure out like for like student scores or something of that sort.
>
> M: Okay. So analysis of data and it's – comes from randomly, random sampling.
>
> I: Um-hum.
>
> M: And you said something about hypotheses?
>
> I: Yeah, like you're trying to find, um, like, probably have to find a comparison if it's, um, it's like – it asks – you have like a question or something and you're trying to see if it's true or f – true or false, so you're gonna use a hypothesis, like the chi-square hypothesis testing, or the – the sample mean, um, hypothesis also, because see that deals with the two sample means.

Based on her statement, it seems that Interviewee 122 understands hypothesis testing to be a means by which a hypothesis can be shown to be true or false. In addition, she focuses on the fact that the data should be randomly chosen. Her comments are somewhat disconnected and it is not clear whether she understands how these ideas relate to each other.

Finally, interviewee number 29 described statistical hypothesis testing as follows:

> **Excerpt: Interview 29**
> I: I think hypothesis testing is um…a way of like…way of I guess…with a big group of people and you're looking for a certain characteristic…characteristic and you're looking for like I guess how many people have like the average? I don't know if that makes any sense (laughs). But, um, yeah like for example like if you're looking at skulls, scores and stuff like how many people get the middle range. And, like, what's, I guess, the probability of being in that middle range…kind of (laughs).

It seems that Interviewee 29 might have confused confidence intervals with hypothesis testing. Though hypothesis testing does connect to the concept of confidence intervals, the goal of hypothesis testing is not to determine a middle range.

Analysis

Though the interviewees were not very articulate in explaining their understanding of the goal of hypothesis testing, their responses give some insight into what they think statistical hypothesis testing is and what it does. Overall, the interviewees seem to understand hypothesis testing as a way for researchers to test new theories or hypotheses. Based on their responses to this question, however, it is not clear whether the interviewees understand the logic behind statistical hypothesis testing, nor is it clear they understand the various statistical and probabilistic concepts that support this logic. It is also not clear whether the interviewees have a strong understanding of the uses of statistical hypothesis testing and/or how to interpret the results within the context in which it is employed. For more insight into these issues, we turn to a presentation and analysis of the data with respect to the identified research sub-questions.

Research Sub-Question Number 2

Results from the multiple-choice assessment indicated that introductory statistics students do not have strong understandings of the logic and reasoning of statistical hypothesis testing. In particular, the quantitative data indicated introductory statistics students do not have strong conceptual understandings of sampling distributions and *p*-values. Furthermore, the data indicated that introductory statistics students do not understand the role of sampling distributions in the overall logic and reasoning of statistical hypothesis testing, nor do they understand the role of inference. When introductory statistics students are told the conclusion of a statistical hypothesis test, they often draw invalid inferences from that information. They often think that a sample statistic provides a direct measure of the population parameter and/or they think that hypothesis tests provide measure of the degree to which the null (or alternative) hypothesis is true. Given these difficulties, however, it was found that introductory statistics students are able to read a scenario and state the correct null and alternative hypotheses for the situation.

As is the case with any multiple-choice assessment, the results of the multiple-choice instrument used in the quantitative phase of this study are limited. They only tell part of the story. By asking the interviewees about their reasoning for the items, it was possible to gain more insight into how students thought about these concepts and ideas.

Six multiple-choice items were chosen to address the second research sub-question in the follow-up interviews and these six items represented five categories of the Framework (see Table 5.1). Analysis of the data will be presented as organized by the five Framework categories. However, given that similar concepts are assessed in the **Analysis of the Sample** and **Decision Rule** categories and for the **Recognizing Applicability** and **Conclusion** categories, data and analysis for these categories will be presented together.

For each section, a summary of student explanations will be presented, followed by an analysis of that data. After the data and analysis for each group of items has been presented, the information will be synthesized into a final summary of the data and conclusions associated with research sub-question number 2.

Generating Statistical Hypotheses

Items in the **Generating Statistical Hypotheses** category were designed to assess whether students understood that, in order to employ indirect reasoning, a researcher must establish the null and alternative hypotheses so that they both specify the measure of the population to be tested and so that they both contradict each other. In the quantitative phase, performance on items from this category was relatively strong. Given a scenario, students were generally able to choose the option that stated the correct null and alternative hypotheses. However, it was hypothesized that this phenomenon may be due to *procedural fluency*, rather than a deep understanding of the logic and reasoning that support the construction of those hypotheses. Introductory statistics students may not have strong understandings of the reason the hypotheses are set up in the way that they are. The follow-up interviews provided an opportunity to explore the degree to which this was the case.

Item number 4 from the **Generating Statistical Hypotheses** category was included in the follow-up interview. This item assessed whether, given a scenario, students could identify the null and alternative hypotheses to be used in a statistical hypothesis test.

Figure 5.1

Item Number 4, Multiple-Choice Assessment

4. A researcher would like to establish that freshmen in the humanities have higher SAT scores than freshmen in the sciences. Which of the following ***null hypotheses*** should be tested?

The mean SAT score of humanities students is …

a. greater than that of science students.
b. greater than or equal to that of science students.
c. less than or equal to that of science students.
d. less than that of science students.

Answer choice *c* is correct. Option *a* represents the *alternative* hypothesis. Options *b* and *d* completed the null hypothesis using different combinations of equality and inequality. Any of the distractors might be chosen if a student does not understand how to state the hypotheses in this scenario.

Overall student performance on this item was relatively high with 32% of the participants choosing the correct answer. Answer choice *a*, however, was the most popular and was chosen by 51% of the participants. These results provided evidence that relative to the other concepts assessed on the multiple-choice instrument, introductory statistics students can successfully establish the null and alternative hypotheses associated with a given situation.

Within the group of interviewees, 3 people chose *a*, 2 people chose *b*, 6 people chose *c*, and no one chose *d*. During the course of the interview, however, 2 interviewees switched their answer from *a* to *c* and 1 interviewee switched from *b* to *c*. Ultimately, by the end of the interview, 1 person had picked *a*, 1 picked *b*, 9 picked *c*, and no one picked *d*.

Interviewee Explanations: A Summary

As the interviewees explained their reasoning about their answer choices, it became clear that there were some commonalities in their thinking. Ten of the 11 interviewees explained that the null hypothesis always expresses equality and that it must contradict the alternative. The rule "the null is always equal" seems to have been strongly reinforced in class. Many of the interviewees claimed that when they set up the problems in class, the null always expressed equality.

Excerpt: Interview 77

I: Yeah. I guess from what we've learned in class like we never really worked with like the null is…like mu is greater than or less than. It's always the mu is always equal to something and then the alternative from that is mu is greater than that claim, or less than that claim [M: Right], or not equal to…

Excerpt: Interview 15
I: I think this one was a little hard because usually in class you do…like the null hypothesis always equals something, but not with a greater than or less than sign with it. Um…so I picked that one because you want to prove that they have higher SAT scores. So the thing that you wouldn't want to happen is that they have equal to scores or just like what you could test and what you really don't want to happen is that they have less than the science scores.

Excerpt: Interview 172
I: Ok. So…for this one…um…I mean in our classes we've been using for the null hypothesis, the mean is equal to a certain number. Um…so…that's why I didn't choose, I guess, *a* or *d*. Um…and…I remember when we were first introduced the subject of, like, null hypothesis that…it was either a greater than or a less than is what some people used and then but the book, our book used equal to so that's what we used in the class…so. I think I just picked that I thought it was greater than or equal to, but…I wasn't sure. But, that's at least how I came to *b* and *c* instead of *a* and *d*.

In addition, some interviewees mentioned that their instructors had given the "null is equal" rule.

Excerpt: Interview 132
M: Why the alternative's not the one that has the equal in it?

I: Ah…yeah I don't know why the alternative doesn't really have equals. I guess cause we were explained, null always has equals (laughs).

Excerpt: Interview 192
M: Why don't you have the testing hypothesis…like isn't it that the alternative is the equal one and the null is the greater than or the less than?

I: I guess because that's the way I've been taught (both laugh). Whenever we do it the…the null is always equal to and then the…whatever that you, whatever…ah…inequality that you want to know. So if you want to know if it's greater than or less than or different that's the…the alternative, the H1.

The "null is equal" rule helped these interviewees to eliminate answer choices *a* and *d*. However, it did not help them to decide between answer choices *b* and *c*. Therefore, more explanation was needed as to how the interviewees chose between these two options.

Some of the interviewees acknowledged that they were not used to seeing an inequality in the statement of the null hypothesis. However, the interviewees reasoned that, since the alternative hypothesis represents the claim (freshmen in the humanities have *higher* SAT scores than freshmen in the sciences) then, though it has an "equals" in it, answer choice *b* cannot represent the null. The null must contradict the alternative. Therefore, answer choice *c* is the better of the two options.

This reasoning is evident in the quotes from Interviewees 15 and 192 presented above, and is also reflected in the following quote from Interviewee number 191.

> **Excerpt: Interview 191**
> I: Ok. Ok, um, I think, looking at these answers, um…it asks, um, the null hypothesis of ah…if freshmen in the humanities have higher SAT scores than freshmen in the sciences. So…I think figuring out the null hypothesis, um, I always assume the null hypothesis to be equal, um, and then the alternative, the hypothesis to have a greater than, or less than, or not equal to, um…so I think a few of these would work. It just depends on which one you would establish it as, whether it's a null hypothesis or alternative, um…So for *c*. *c* is the answer I chose. Mean SAT scores of humanities students is less than or equal to that of science students. Um…I think I chose that as opposed to, ah, *b*, which is greater than or equal to that of science students, um, because I think looking back at it, I think the greater than value would be, um, the alternative hypothesis, because that's what the researcher's trying to prove or establish, and that's why I chose *c* as the null hypothesis.

Though the interviewees understood that the null hypothesis is a statement of equality and that it must contradict the alternative hypothesis, they did not have a good understanding of why this is the case. As demonstrated in the excerpts from the conversations with Interviewees 132 and 192 above, when the interviewees were asked to explain why the null hypothesis must be a statement of equality that contradicts the alternative hypothesis, many of them cited a rule about the null hypothesis being equal and/or about the alternative hypothesis representing the claim of the researcher. In fact, when asked whether the two could be switched (the alternative hypothesis claims equality while the null claims inequality), many of the interviewees thought the two could be switched.

> **Excerpt: Interview 81**
> I: Um…I guess if you switched them…it probably wouldn't…um…but I'm really not sure…as long as you switched them evenly.
>
> **Excerpt: Interview 191**
> I: Well, I'm just, again, it's like with this, um, they're just opposites, so it depends on which one you want to be, your null, or alternative. Either way, you'd still prove, you could still prove the same answer. It's just that it, like for this one, um…I chose *c* as my null hypothesis. If *c* was, or if less than was wrong, um, then that means my null hypothesis would be wrong. If I reverse these two and I chose greater than, um, as my null hypothesis, and that was right, that would still prove the same information.
>
> M: Ok, so it doesn't matter.
>
> I: Yeah, it I don't think that's as important as long as you, you'd, like these, um, are just complementary. The data would still prove the same answer regardless of whether this was your null or that was your alternative.
>
> M: Does there have to be an equal in there somewhere? Like for example, could the null have been less than and the alternative be greater, so a researcher wants to establish that the freshmen in the humanities have higher SAT scores, so that would have – the researcher – maybe alternative being that the humanities is greater than the science, and the null that the humanities is less than the science, so there's no equal in either of them. Is that possible do you think?
>
> I: Yeah. Um, again, I guess it depends on, ah, what question you wanna answer.
>
> **Excerpt: Interview 169**
> M: So why isn't it…could I have it something that I'm trying to prove that the null…or, sorry that the mean is equal to 15 and the null is that the mean is less than 15?
>
> I: Say it again…I just…I thought I heard you…
>
> M: I know…so let's say I want to prove that means are equal…and it's assumed that the one mean is less than the other. So that would be my null. Could I do that?
>
> I: I don't know if I've seen that done…but, I mean, I suppose it's reasonable if you…yeah, I've, I've never seen that done. But, again, it seems like it's possible.

These responses indicate a weak understanding of the logic and reasoning of statistical hypothesis testing. Statistical hypothesis tests rely on indirect reasoning to draw inferences about populations based on information from a sample. To do so, the null hypothesis is assumed to be the true descriptor of the population and the sample is analyzed to determine the degree to which it is unusual under the assumed null condition. Sampling distributions are used to find the probability of obtaining a test statistic as extreme, or more extreme, than the observed under the assumed null condition. If the probability is small, the sample is deemed unusual and the null hypothesis is rejected. In order to employ this logic, the null hypothesis must describe a clearly defined population. Then, a sampling distribution can be defined and used to determine if the sample collected is unusual. Therefore, the null and the alternative hypotheses can not be switched. The null must be the assumed hypothesis. The interviewees in this study did not recognize this to be the case.

Analysis

As was hypothesized in the quantitative phase, the interviewees' reasoning about this item was "rule based". The interviewees remembered that the alterative hypothesis should represent the claim that the researcher is attempting to demonstrate and that the null hypothesis contains a statement of equality that contradicts the alternative hypothesis. However, the interviewees did not have a strong understanding of why those "rules" exist. They did not understand that statistical hypothesis testing relies on indirect reasoning and assesses the degree to which the sample presents an unusual case conditioned on the null hypothesis. Interviewee 112 came closest to expressing an understanding of this logic. He understood the logic of proof by contradiction. However, based on his comments, it was not clear that he understood that statistical hypothesis testing does not rely on proof by contradiction in the formal sense, but that it relies on probability to determine whether a sample is *unusual*. It is not clear that Interviewee 112 understood this logic and, based on their responses to item number 4, it certainly wasn't clear that any of the other interviewees had this understanding.

Analysis of the Sample and Decision Rule Categories

Items from the **Analysis of the Sample** and **Decision Rule** categories were written to assess whether introductory statistics students understand the concepts and reasoning involved in deciding whether or not the null hypothesis should be rejected. These items address sampling distributions and their role in determining whether or not a sample is unusual, conditioned on the null hypothesis. The items assess whether students understand that, if a sample is deemed unusual, the result is said to be statistically significant and whether they understand that the level of significance (decision rule) has an impact on that result. These items also address the concept of *p*-value and its role in the progression from sample analysis to the statement of a conclusion.

In the multiple-choice assessment, performance on items in the **Analysis of the Sample** and **Decision Rule** categories was relatively weak. The concepts of sampling distribution, *p*-value, and significance level were troubling for students. The quantitative results indicated that introductory statistics students do not understand that a sampling distribution represents the distribution of sample statistics calculated for all possible samples of a given size from a population with a related, known population parameter. The results also indicated that introductory statistics students do not understand that the level of significance signifies the probability with which the researcher is willing to make a Type I error and that it provides the "cut point" for which a sample is deemed unusual. Finally, the results indicated that introductory statistics students do not understand that the sampling distribution and decision rule are useful in deciding whether or not to reject the null hypothesis.

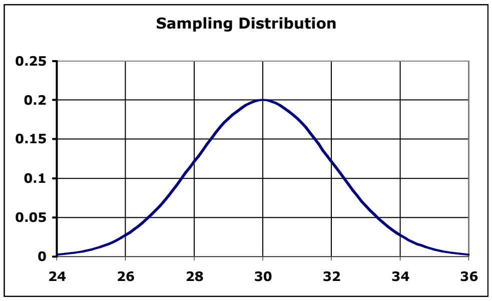
However, the results of the quantitative phase tell only part of the story. The follow-up interview included one item each from the **Analysis of the Sample** and **Decision Rule** categories and provided more insight into student thinking about these components of statistical hypothesis testing.

Item number 9 was included in the follow-up interview. It is classified in the **Analysis of the Sample** category from the Framework and was written to assess whether introductory statistics students understand how sampling distributions are used in determining whether a sample is unusual conditioned on the null hypothesis.

Figure 5.2

Item Number 9, Multiple-Choice Assessment

9. Tests show that fuel efficiency of cars in the current model year averages 30 miles per gallon. A test of 100 new car models gave mean fuel efficiency of 31.5 miles per gallon. To see whether it is correct to claim that the new car models are more fuel-efficient than those in the current model year, a researcher constructed the sampling distribution of average fuel efficiencies of many samples of 100 current model cars shown in the following graph.



Which of these conclusions is ***best supported*** by the graph above?

a. The difference in fuel efficiency of current and new model cars is not statistically significant.

b. Half of current and new model cars have fuel-efficiencies below 30 miles per gallon.

c. The difference in fuel efficiency of current and new model cars is statistically significant.

d. Nothing can be concluded. The graph should be centered at 31.5.

The correct answer is option *a*. Answer choice *c* is the opposite of *a*, and answer choices *b* and *d* are based on misunderstandings identified in the literature.

On the multiple-choice assessement, 40.4% of the participants chose *a*, 21.2% chose *b*, 8.7% chose *c*, and 27.9% chose *d*. These results indicated that while many introductory statistics students were able to read the graph to determine whether the results were statistically significant, a good number of students are not able to do so.

Of the eleven Interviewees, 6 chose *a*, 2 chose *b*, and 2 chose *c*, and 1 chose *d*. Over the course of the interview, however, the interviewee who chose option *d* changed her answer to option *c*. Unfortunately, the data collected in the interviews indicated that though over half of the interviewees chose the correct answer, they did not have strong understandings of sampling distributions and their role in determining whether or not a result statistically significant.

Interviewee Explanations: A Summary

In the analysis of the data, it was abundantly clear that the interviewees did not know what the graph associated with this item represented. The interviewees were not familiar with the term "sampling distribution" and did not realize that the graph represented the distribution of average fuel efficiencies for all possible samples of size 100 in a population where the average fuel efficiency is 30 miles per gallon. Five interviewees (15, 81, 122, 169, and 191) were not able to interpret the graph at all. Five of the remaining six interviewees (29, 77, 112, 172, and 192) understood that the *x*-axis referred to fuel efficiency. However, it was not clear that these interviewees understood that the *x*-axis referred to *average* fuel efficiency in a sample of 100 current cars. Of those five, Interviewees 112 and 192 thought the *y*-axis represented the percentage of cars in "the sample" that had a given fuel efficiency. Interviewee 77 thought the *y*-axis represented the probability that a car in "the sample" had a given fuel efficiency. Interviewee 29 thought the *y*-axis represented the proportion of cars in the population that had a given fuel efficiency. Interviewee 172 thought the *y*-axis represented the standard deviation. Interviewee 132 had a completely different interpretation of the graph. She thought the *x*-axis represented the number of cars out of "the sample" of 100 and the *y*-axis represented the number of cars out of "the sample" of 100 while the *y*-axis represented the difference from the mean (of 30).

The phrase "the sample" here is in quotes because, at times, the interviewees confused the type of car represented in the sample. Though the interviewees could read the item and understood it to tell them that the graph represented the sample means of samples of *current* cars, the interviewees would often reason as if the sample of 100 cars had been drawn from the population of *new* cars.

It is interesting to note that though they did not understand precisely what the graph represented, none of the interviewees chose option *d*. They reasoned that (1) something *could* be concluded and/or (2) the graph should be centered at 30 because, in the text, the reader is told that the graph represents the sampling distribution of *current* cars. With answer choice *d* eliminated, the interviewees focused on answer choices *a*, *b*, and *c*.

The way that the interviewees interpreted the graph influenced the reasoning they used to justify their choice of *a*, *b*, or *c*. Interviewees 15 and 169 chose *b* because they did not think that graphs could be used to determine statistical significance.

Excerpt: Interview 15
I: Well for that, cause like the graph just looks like it's a distribution…like you just took all your data from your current cars and you put it in a graph. So it's not like…um…there's no mention of like an alpha or like a confidence interval or anything or a test. So you can't like say that it's statistically significant like you could hypothesize that like oh maybe it's significant but you're not like analyzing it at all you're just kind of like here's a graph kind of thing.

Excerpt: Interview 169
M: *B* Sorry, yeah. So *a* and *c* are kind of opposites, whether it's statistically significant or not [I: Right]. So why didn't you pick either of those?

I: Well, it's not supported by the graph. The graph…as far as I'm…as far as I think, I don't think a graph can tell me about statistical significance. A test statistic has to…and, ah, nowhere on the graph above does it make a comparison of a current versus a new model. There's just a simple line…so.

M: Ok, so the fact that the second thing gave me 31.5…and the other one's 30…so the, the means in the [I: Right] in that was different. So…that makes them…how would

> they be statistically significant? I mean, I could say well they must be different because I got 31.5 [I: Right] and it was 30.
>
> I: Um…well you'd have to use a hypothesis…you'd have to use a test statistic and then a rejection region to determine if it was statistically significant…meaning, I guess statis, statistically significant to me means…if you accept H1 and reject H0.

Because Interviewees 15 and 169 did not know how to interpret the graph, and because they did not think a graph could be used to determine whether or not a result is statistically significant, they chose option *b*. That answer choice did not mention statistical significance.

Interviewees 81 and 191 also did not know how to interpret the graph. However, they chose the correct answer, option *a*. Interviewee 81 did so because he didn't think there was a big difference between a mean of 30 and a mean of 31.5.

> **Excerpt: Interview 81**
> I: Yeah, I chose *a*…the difference in fuel efficiency of current and new model cars is not statistically significant. Um…I figure that was the best answer because the difference between the two is very slim and you're only testing 100 cars. You know there's…um…more than 100 new car models out there. Um…and *b* said half of current and new model cars have fuel-efficiencies below 30 miles per gallon. Um…let me see…well first off…I wasn't really sure what this graph was saying because it wasn't really labeled. Um…so I, I didn't really like that and then *c* the difference in fuel efficiency of current and new model cars is statistically significant. Well, I felt it wasn't statistically significant so obviously saying it was is the opposite.

Interviewee 191 chose *a* for a different reason. His reasoning is illustrated in the excerpt included below:

> **Excerpt: Interview 191**
> I: Um, so between *a* and *c*, not statistically different. I – I think it says that there was, um – well, this – the graph shows that there's many samples of 100 current model cars, um, that were taken, um, so many samples of 100, whereas, um, with the, ah, test of new cars, there was only 100 that were sampled or tested, um, so I think, ah, ah – I'm trying to think. Although I guess the test would be better if they did many samples of 100 new car models, um, I remember we went over the, ah, central limit theorem, and I think, ah, that with *n* being greater than 30, that you can assume that it's normal, and that I guess it's significant. Um, so I think that the answer would be *c* in that case.
>
> M: And so why are you choosing *c* now?

> I: Um, because they tested 100 new cars, although they tested – that it was obviously many more than 100 cars, current cars, um, the fact that they tested 100 new cars, even though they tested many more current cars, um, because 100 is greater than 30, you can use the central limit theorem, um, to assume something about the population, um, and because 31.5, the fuel efficiency they determined of the 100 new cars, um, is greater than 30, um, it's statistically significant.

Interviewee 191 reasoned that because the sample was greater than 30, one can assume it is "normal". Therefore, the result is statistically significant. For him, statistical significance means that "it can be approximated to the normal curve". Thus, Interviewee 191 did not use the graph to answer the question. He merely applied some (incorrect) definitions he remembered from class. Like Interviewee 81, Interviewee 191 did not use the graph in his reasoning, but chose the correct answer. In both cases, however, the reasoning used was not correct.

Interviewees 29, 132, 112, and 192 also chose the correct answer, option *a*. However, unlike Interviewees 81 and 191, Interviewees 29, 132, 112, and 192 *did* refer to the graph in their reasoning. Unfortunately, these interviewees did not interpret the graph correctly in the first place. Their misinterpretation of the graph caused them to rely on incorrect reasoning in choosing the correct answer.

Interviewees 112 and 192 both thought the graph represented the percent of cars in the sample that had a given fuel efficiency. Interviewee 112 relied on this interpretation to justify his answer choice.

> **Excerpt: Interview 112**
> I: I think I put *a* because it looked like…it pretty much looked like how the old one would be also…like the old model averages 30 and this was centered around 30 also. And, *d*, nothing can be concluded. The graph should be centered at 3 point…31.5. I was thinking maybe outliers made that…and be like 30 if you looked at all the data…it would be like a more accurate measure of center. But as it's almost the same, it's not statistically significant, which is why *c* is wrong. And, I also think …think you could assume that's half and half. So probably is outliers in there.
>
> *discussion continues*

M: And why did you pick *a* again? I know I wrote it down but I want to just…now that I know what you're thinking about with the graph…why did you pick *a*?

I: Ok, cause it looked like that since most were …like 20% were around what the old average miles per gallon…and how it seemed like there were around 50% of each were going the other way, that it wasn't really statistically significant. Probably that the mean of 31.5 miles per gallon was just derived using outliers. Cause I was assuming it was a correct graph, unlike what *d* said that it should be centered at 31.5. Although I could be wrong about that.

Interviewee 112 interpreted the point (30, .20) to signify that 20% of the cars had fuel efficiency of 30 miles per gallon. Therefore, the difference wasn't statistically significant. In addition, Interviewee 112 reasoned that a mean of 31.5 was the result of having outliers in the sample. If these outliers had not been included, the sample would have a mean of 30 as indicated by the graph. He also thought 31.5 was not too different from 30. Although his reasoning wasn't entirely clear, Interviewee 112 did not think the difference was statistically significant and chose answer choice *a*. He chose the correct answer for the wrong reason.

Interviewee 192 interpreted the graph in the same way that Interviewee 112 did. However, she chose answer choice *a* for a very different reason. Her reasoning is illustrated in the following excerpt.

Excerpt: Interview 192
I: It just, yeah, it just…it…had the normal shape. [M: Ok] I did realize something that it was wrong with the way it was centered. So…I just chose *a*…cause I know that you could conclude…I figured that you could conclude something from that. So that's why I didn't choose *d*. And, I couldn't really read it, so I wasn't sure about *b*. Even though I couldn't read it…regularly.

discussion continues

M: Ok, and *a* you chose because since it looks normal [I: Yeah] but not correctly normal [I: Yeah]…(writing) but not correctly normal, it's not statistically significant. And if it were centered with the y-axis in the middle, then you would have picked *c*?

I: Yeah…

Interviewee 192 associated the normal curve with statistical significance. However, for Interviewee 192, a result is only statistically significant if that normal curve is symmetric about the *y*-axis. Because the graph in item number 9 is not centered over the *y*-axis, the result is not statistically significant. Therefore, Interviewee 192 chose option *a*. Unfortunately, she chose the correct answer for the wrong reason.

Interviewee 132 also employed incorrect reasoning in her choice of the correct answer, option *a*. Her interpretation of the graph influenced her reasoning. Her thinking is illustrated in the following excerpt.

> **Excerpt: Interview 132**
> M: Ok. So now that I know how you're reading the graph…[I: Ok]…what, you chose *a*….[I: *a*]…why?
>
> I: Ah…well looking at it, it looks like, ah, well 24 and 36 of the cars are at same level. So that's 60 cars, and so that means 60% of all the new cars have the same av…as the old gas mileage. So, it doesn't look like they're significantly different. Cause the 60 just seems too high…it's more than half.

Here, Interviewee 132 focuses on the points (24, 0) and (36, 0) on the graph. Because, for her, the *y*-axis represents difference from 30, she interprets these points to signify that 24 + 36 = 60 cars have a fuel efficiency of 30 miles per gallon. That means that 60/100 = 60% of the new cars have the same mean as the older cars. Because 60% is more than half, the difference is not statistically significant. Using this incorrect reasoning, Interviewee 132 chose the correct answer, *a*.

In her explanation for choosing answer choice *a*, Interviewee 29 came the closest to using correct reasoning. Her thinking is illustrated in the following excerpt.

> **Excerpt: Interview 29**
> I: Ok, I was confused on this one, so I kind of guessed (laughs). Cause I wasn't sure what like…well…no, never mind. I, I picked this one I think because it said the mean was 35…or 31.5 and like…and the distribution like the middle range was like around 30…so that's why I picked *a* because it didn't look like it was very different.

Her explanation indicates that she was looking at the graph to see how close 31.5 was to 30. She thought they were close on the graph, so she did not think that the difference was statistically significant. This reasoning is correct. However, given that Interviewee 29 thought the graph represented that proportion of cars in the population that have a given fuel efficiency (rather than a distribution of *average* fuel efficiencies in *samples* of size 100), the degree to which she has a deep understanding of the way in which statistical significance is determined is suspect.

Of the three interviewees who chose option *c*, only one referred to the graph. Like Interviewee 191, Interviewee 172 focused on the fact that the graph had the shape of a normal curve.

> **Excerpt: Interview 172**
> I: Ok. All right. Um…ok, I think…well…I remember this one…I like went back and forth about this one but I think I chose *c* just because it resembles a bell shape curve so it could be approximated to the normal distribution which could, I guess, make it significant in terms of…um…if you wanted to do any type of hypothesis testing with it you knew that it could be approximated…um… and I didn't think *a* because I…, I mean it…I definitely think it is statistically significant because it resembles a bell shaped curve. And, then…um…(discussion about flipping the tape). Um…so…right, I didn't, I didn't choose *a* because I did think it was signif, statistically significant because it resembled a bell shaped curve.

Here, Interviewee 172 saw that graph was "normal" which meant that it was statistically significant. Unlike, Interviewee 191, she was not concerned with the fact that it was centered at 30 rather than over the *y*-axis.

Interviewees 77 and 122 did not use the graph in their reasoning for choosing option *c*. Instead, they focused on the fact that there was a difference between the mean of the sample and the mean of the current cars.

> **Excerpt: Interview 122**
> I: Hum, I think *c* can work. Um, the significance and fuel efficiency of current and new model cars assumed that this statistically significant because there is a difference between the average from the current model then new model also. So there's the change.

> **Excerpt: Interview 77**
> I: Ok. Well, it's not *d* because…because it's a graph of the current model cars and not the new ones. The average is 30 miles per gallon, not 31.5. And, it's centered at 30 so that's right. And, um…I said that…the difference in the fuel efficiency of the current and new model cars is statistically significant because there is a difference in the mean fuel efficiency that the cars get and…I didn't really take this from the graph…because, the graph just showed me that this was a…a distribution of the current model cars, and their fuel efficiency, which is 30…um. But, from what the problem said that, the new car models…had a mean fuel efficiency of 31.5 compared to the current car models with a mean fuel efficiency of 30 miles per gallon. Since there was a difference, I said that was significant.

Like Interviewee 81, Interviewees 122 and 77 focused solely on the difference between 31.5 and 30. Interviewee 81, however, had some understanding that statistical significance should take into account the size of the difference with respect to the size of the sample. As a result, he did not think the difference in means was significant. Interviewees 77 and 122 did not take sample size into consideration. They considered *any* difference in means to be statistically significant.

The discussion above provides evidence that the interviewees did not have deep, connected understandings of sampling distributions and statistical significance, and how the two relate. They are unfamiliar with the term "sampling distribution" and, on the whole, do not have strong understandings of what it means for a result to be statistically significant. Even when they chose the correct answer, the interviewees used the wrong reasoning. Overall, the group of interviewees did not have strong understandings of these concepts and ideas.

Tied into the concept of statistical significance is the *level* of significance, or the decision rule. Examination of student thinking about the item representing the **Decision Rule** category provides additional insight into student thinking about the way in which statistical hypothesis testing can be used to draw a conclusion about the null (and alternative) hypotheses.

Item number 11 was included in the follow-up interview. This item is classified in the **Decision Rule** category and assesses whether introductory students understand significance level and its role in hypothesis testing. In particular, this item assesses student understanding of the effect of different significance levels on the overall conclusion of the statistical hypothesis test.

Figure 5.3

Item Number 11, Multiple-Choice Assessment

11. To test the hypothesis that private schools are more effective than public schools, researchers plan to compare mean starting salaries of private and public school graduates. But they cannot agree on whether to test the results at a significance level of 0.10 ($\alpha = 0.10$) or at a significance level of 0.05 ($\alpha = 0.05$).

What effect will using 0.10 rather than 0.05 have on the study?

a. Using 0.10 will result in a greater chance that they will incorrectly retain the null hypothesis.
b. Using 0.10 will result in a greater chance that the null hypothesis is actually true.
c. Using 0.10 will result in a greater chance that they will incorrectly reject the null hypothesis.
d. Using 0.10 will result in a greater chance that the alternative hypothesis is actually true.

The correct answer is option *c*. Answer choice *a* states the opposite. Answer choices *b* and *d* are counterparts to *a* and *c*, except they make general claims about the truth of the hypotheses.

In the multiple-choice assessment, 33.7% of the participants chose option *a*, 28.8% chose option *b*, 29.8% chose option *c*, and 7.7% chose option *d*. These results indicate that introductory statistics students do not have strong understandings of the effect of significance level in a statistical hypothesis test.

Of the 11 interviewees, 5 chose *a*, 1 chose *b*, 4 chose *c*, and 1 chose *d*. However, over the course of the interview, 3 people changed their answer from *a* to *c* and 1 person from *d* to *c*. Therefore, after the interviews were completed, 2 people had picked *a*, 1 person picked *b*, 8 people picked *c*, and no one picked *d*.

Interviewee Explanations: A Summary

With the exception of one person, the interviewees did not choose answer choices *b* or *d*. It was hoped that the interviewees eliminated these options because they each make very broad claims about the truth of the either the null or alternative hypotheses. This was the case for two of the interviewees.

> **Excerpt: Interview 15**
> I: Um…I guess because like when you're doing the test you're kind of between like either accept or reject and not necessarily true or false.
>
> M: Why don't you…do you know why you don't do true/false? Why they keep telling you to use those words accept and reject?
>
> I: I guess cause like…like it seems to me like you can't really prove anything cause true is really concrete and you can't prove anything absolutely unless you go and take like every bit of data available to you. And, you …like in most cases you actually can't physically do that. And, like even where you can, like it's just really difficult to do so it's not practical, I guess. Like …it's very rare, I suppose that you'll come across a study when they can say like this is true because I went out and sampled every single student at every single college and, found that this would be true. But, then like, again, it would only be true for like that point in time, at that instant. Like, you can't really say something like so concrete about like such a…like a big thing.
>
> M: Ok.
>
> I: So like if you say reject, you're saying like you're not talking absolutes. You're talking more of like in relative terms for what you worked on.
>
> **Excerpt: Interview 112**
> M: …why not *b* or *d*? Feel free to…if you need to write at any point…
>
> I: (reads *b* out loud, somewhat inaudible) They pretty much are saying the same thing (laughs)
>
> M: Ok

I: Well, I guess you can't ever know for sure if it's true, though. That's probably why I didn't look at those.

For other interviewees, it was not clear whether they had deep understandings of why these options are incorrect. The following excerpts illustrate these somewhat incomplete explanations.

Excerpt: Interview 169
I: Well…(inaudible) I mean, *b* and *d*…I don't believe that you're…and maybe there is some correlation between your significance level and whether or not your null hypothesis has a greater chance of being true…or of either hypothesis actually being true. But, I, I didn't know, and still don't know that answer. So I can't…those, those answers don't seem right to me.

Excerpt: Interview 132
I: …But…ah…I guess *c* would be better than *d*.

M: Why?

I: Um…because you're saying that the…you could incorrectly reject the null hypothesis whereas in *d* they talk about the alternate hypothesis being actually true. So…just your…your acceptance or rejection region doesn't determine if something is true or false. It just is…it's naming the, your parameters for you graph to label acceptance and rejection.

Excerpt: Interview 29
I: Well, *b* it's not…like 0.1 would then give you a greater chance that null hypothesis is actually true…0.1 gives you a greater chance that it's rejected.

Still others did not pick *b* and/or *d* because they did not know whether the null hypothesis was actually rejected or not. This was the case for Interviewees 77 and 172.

Excerpt: Interview 77
M: …Why not *b* or *d*?

I: Um…well because if you use alpha 0.10…with the *p*-value like I said 0.8, you're rejecting null hypothesis…and that's not true. Cause if you reject the null hypothesis then the null hypothesis is…is…isn't actually true. Um…and with that said, I guess *d* could be possible. Cause if the null hypothesis isn't true then the alternative hypothesis is. But, based on what I learned with using alpha, that use…that's most…mostly…you usually use that for the null hypothesis, not so much with the alternative hypothesis. Because…well, yeah. Yeah.

Excerpt: Interview 172
I: Right, because…it's the opposite. And, then, ah…I don't, I just, I didn't think *b* and *d* were even like correct because you didn't really know anything about how it turned

out. Like, if it was rejected or accepted, though I think they're talking about is they just can't decide on what significance level to test it at.

M: So because the, you don't know whether they accept or reject, you don't know whether *b* or *d*?

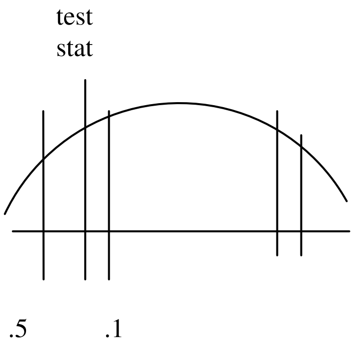
I: Right.

In both cases, the interviewees made their decision about *b* and/or *d* based on whether or not they thought the null hypothesis was rejected or not. This reasoning indicates that Interviewees 77 and 172 did not reject *b* and/or *d* because the claims were too broad.

With the exception of Interviewee 81 who simply guessed when he chose answer choice *b*, the interviewees chose either *a* or *c*. Some of the interviewees justified their choice by referring to rejection regions on a graph, some used numerical justifications, and others simply expressed "rules" to justify their choice.

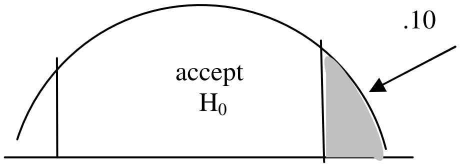
Interviewees 191, 29, and 15 referred to graphs as illustrated in the following excerpts. The accompanying graphs were drawn by the interviewees to support their explanations.

Excerpt: Interview 15
I: Um…this one…like the…when you have, um, a bigger significance level, that means like you're more sure what your result can be. That's like…in confidence intervals, at least… like you're more sure that the mean is in there but it doesn't necessarily mean you're more precise. So with, like a 0.1 confidence interval instead of 0.5, like if you look at the picture, the…the um… confidence intervals are like… (draws)…like a 0.5 would be out here and then…um…because you'd have a greater area your 0.1 would be out here. And, like if you did a test, then your test statistic would be right there then if you use the .5 then you're accepting the null hypothesis. But, if you use the 0.1, then you're rejecting it. So…um…there's a greater chance that you might incorrectly reject it.



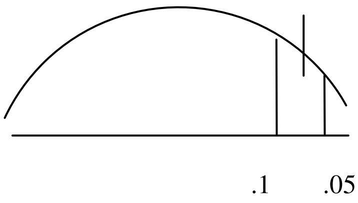
Excerpt: Interview 29

I: (Reads *a* somewhat inaudibly) Well…it also says like will result in a greater chance that….they will incorrectly retain the null hypothesis. Well, 0.10 is like …it's usually the area at the ends, which is like…um…the rejection region for the null hypothesis. So…yeah…so I said…um…it will not retain…I don't know…(laughs). You know, because 0.1 is the rejection region for like the null hypothesis and that's why I didn't pick that one…(inaudible).



Excerpt: Interview 191

I: Well, 'cause – Ok, so if this is a one sided test, if alpha is this, if that's .05 and this is .01, um, if your value that you calculated fell here, um, um, point – what was that… yeah,… .10 means that there's a greater chance, um, that yeah, so if this was the value that proved that your null hypothesis was right – if you use .05, then that means it's – wait a minute. No, it would be *c*, I think, ah, because you would reject – if the null hypothesis fell in that region, you would reject it at .1, because it's beyond that value. If it was at .5, it would still be within that value, so the answer is *c*.



Interviewees 191 and 15 reasoned that if the test statistic fell in between 0.05 and 0.1, then the null hypothesis would be rejected at the significance level of 0.05 but not at 0.10.

Therefore, *c* must be correct. Though Interviewee 29 also used a graph, her reasoning was slightly different. She reasoned that since the graph shows *rejection* regions, then *c* must be correct, not *a*. The claim in option *c* contains the word "reject" while the claim in option *a* contains the word "retain".

It is interesting to note that, though the discussion surrounding item number 9 indicated that Interviewees 15, 29, and 191 did not understand graphs of sampling distributions, these interviewees used graphs to justify their answer choice for item number 11. As a result, the degree to which they understand why the graph they drew justifies their answer to item number 11 is questionable. In fact, further conversation with Interviewee 15 revealed that, indeed, she didn't understand why the graph supported her answer.

Excerpt: Interview 15

M: So what is this graph of that you drew? I mean, I understand the point…but what does this curve represent?

I: Um…like it would be either like the *z* or the *t* depending on like how big your sample size is.

M: *z* or *t* distributions?

I : Yeah.

M: What are those things? Do you know?

I: No.

These comments provide evidence that Interviewee 15 did not have a deep understanding of the way in which the graph supported her answer choice. Due to time constraints, the same questions were not asked of Interviewees 29 and 191. Therefore, the question as to the depth of these interviewees' understandings with respect to this issue remains unanswered.

Interviewees 112 and 77 also chose *c*, but they did not refer to graphs to justify their choice. Instead they relied on calculations and *p*-values. The following excerpt illustrates Interviewee 112's reasoning.

Excerpt: Interview 112
I: All right…(reads out loud, somewhat inaudible) I put *c* because 0.10 is…it's less accurate than 0.05 cause there's a more chance that they'll incorrectly reject the null hypothesis. Ok, cause it's much more of a range. Ok cause they're not going to retain the null hypothesis more with…at 0.10. Ok…all…(reads out loud, somewhat inaudible)…wait a minute…I might of…I might have had this backwards, though…(reads out loud, somewhat inaudible)…cause it's gonna be larger with 0.10…yeah, so it might actually be *a*. They might incorrectly retain the null hypothesis when they could have rejected it. Cause it's a larger range with 0.10. Yeah, and it's…they didn't correctly reject the um…the non-null hypothesis, the one they're trying to test. Yeah, so I'll change that one to *a*.

He writes the following on his paper

$$\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \quad , \quad \bar{x} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

$$\left(\bar{x} - 1.58\frac{s}{\sqrt{n}} \quad , \quad \bar{x} + 1.58\frac{s}{\sqrt{n}}\right)$$

$$\begin{array}{l} R : x \geq z_{.10} \\ x \geq 1.58 \\ x \geq z_{.05} \\ x \geq 1.645 \end{array}$$

I: (Writes) The *z*-value of 0.10…it's gonna be larger…so that would be x…five eight…(writing and talking out loud, somewhat inaudible)…I might have had it right to begin with…looking at it this way (laughs). It's easier to reject 1.58 than 1.645.

M: Ok. What was that that you were writing out there? What was all this?

I: Oh, first I was doing like…like the confidence interval. Then I realized that wasn't going to do anything good (laughs).

M: Ok

I: Yeah, so then I did rejection range.

M: So this is different from all of that (draws black line to separate work)

I: Yes. Um hum. This is the stuff that actually matters (points to right side of black line).

M: Ok. And what were you doing? X greater or equal to z…

I: Yeah, I was assuming it was a *z*-value.

M: Ok. And, what's the 1.58?

I: For the…the alpha equals 0.1.

M: And so what does 1.58 represent?

I: The *z*-value for 80% confidence.

After having done a calculation to find the confidence interval, Interviewee 112 changed his approach and, instead, wrote inequalities that represented the requirements that would result in rejection of the null hypothesis under significance levels of 0.10 and then of 0.05. With these calculations, Interviewee 112 reasoned that it was "easier" to reject the null hypothesis under a significance level of 0.1 rather than at a significance level of 0.05. Therefore, he chose answer choice *c*.

Interviewee 77 also relied on a numerical justification for her choice. However, she referred to *p*-values in her explanation. Her reasoning is illustrated in the following excerpt.

Excerpt: Interview 77

I: Ok. Um…well…from what we learned about using alpha levels…um…to find out whether to reject or accept a null hypothesis…um. I knew that when we take the *p*-value…or when we would find the *p*-value of the test statistic, that's the …to determine the probability that …um…that the value is more extreme…um…probability that…yeah…so, the *p*-value is the probability that the value you find is more extreme than the…then the value you're testing? And, once you find that *p*-value, you compare that to the alpha level. And, if it's less than the alpha level, then you reject the null hypothesis. If it's greater than the alpha level or it's not less than the alpha level then you accept the null hypothesis. So if you were to use a alpha level of 0.1 compared to 0.05…um…I said that's true that they…there will be a greater chance that they will incorrectly retain the null hypothesis. Because 0.05 is a, it's a much smaller alpha value…or alpha level…um…and if you use a smaller alpha level…alpha level…it's kind of a stronger, I guess, a stronger number…it kind of…ok, well. Like say you get the *p*-value is like 0.8 and so if this is the *p*-value, then you compare that to alpha. If alpha is 0.10…that would …that's saying that the *p*-value is less than alpha so you reject the null hypothesis. But if you compare that *p*-value, the same one to 0.05…it's not less than the alpha. It's greater. So, you're saying accept the null hypothesis. So, kind of using a smaller alpha value gives you…a more…a certain…decision on what to do with the null hypothesis…um…

She wrote the following on her paper – arranged in precisely this way

> P(you get a more extreme value than with μ_0)
>
> p-value .8 .05
> $\alpha = .10$ accept H_0
> reject H_0
>
> I: Ok, so I think I picked the wrong one because I picked the one that says incorrectly retain, when it should be the one that says incorrectly reject, if you're using alpha 0.10…cause, I just…did not (both laugh).

Interviewee 77 reasoned that if the significance level was set at $\alpha = 0.10$ and the *p*-value was calculated to be 0.08, the null hypothesis would be rejected. However, if the significance level was set at $\alpha = 0.05$ and the *p*-value was calculated to be 0.08, the null hypothesis would be retained. Thus, it was "easier" to reject the null hypothesis under a significance level of 0.10 than at a significance level of 0.05. Therefore, she chose answer choice *c*.

It should also be noted that Interviewee 77 did not state a correct definition for *p*-value. An important piece of defining *p*-value is the notion that, while it is a probability, it is a probability *conditioned* on the null hypothesis. It is not clear whether Interviewee 77 understood that the *p*-value is the probability of a obtaining a test statistic as extreme, or more extreme, than the observed, given that the null hypothesis is the true descriptor of the population.

As was the case for Interviewees 15, 29, and 191, it is not clear that Interviewees 112 and 77 had deep understandings of why their work on item 11 justified their answer. Though Interviewees 112 and 77 used correct reasoning from a procedural standpoint in determining the effect of various levels of significance on the conclusion, it was not clear that they understood that their calculations were based on the assumption that the null is the correct description of the population under question.

Other interviewees reasoned without graphs or calculations and, instead, stated rules about the relationship between significance level, certainty, and size of rejection regions. The following excerpts from Interviewees 132 and 169 illustrated such reasoning.

Excerpt: Interview 132
I: Ok (reads out loud, somewhat inaudibly)…um...well I said *a*…but, um…if you use…ah…the area of significance of 0.1 or 0.5…if you use 0.5, you're acceptance region for the null hypothesis gets larger…if you're using like 5%. Um…so I guess I said *a* because you could…can…you could keep the null hypothesis even if it was very far out to the sides. Um…oh, I mixed it up. Because it's 10% it should be a smaller acceptance region. Can I change my answer now?

discussion continues

I: Oh…um...if like…if has like a 95% confidence interval that means it's going to have like a larger…ah…start, endpoint. So anything in between those two endpoints would be accept…you'd accept your null hypothesis. But, if you made your acceptance region smaller, that means there's more chance that you would reject your null hypothesis.

Excerpt: Interview 169
I: Um…and…when it's asking for a significance level…as I believe, if I understand it correctly…and it's entirely possible that I don't…um…if I…increase the significance level that means I decrease my certainty and that means…for um, the rejection region that I decrease the size of my rejection region…meaning that…there…by the law…by the, you know,…by, you know, intuitive probability that there is a smaller chance that I…accept my…second hypothesis. So…*a* seems to verbalize that best for me.

M: Ok. So increase of significance level means decrease in certainty which mean decrease in the rejection region [I: Yes] which means a smaller chance you'll accept the null. Did I get that right?

I: A smaller chance that you'll accept the secondary hypothesis.

Here, we see that Interviewees 169 and 132 used a rule to justify their choice. Interviewee 132 stated the correct rule: the larger the significance level, the larger the rejection region. Therefore, increasing the significance level from 0.05 to 0.1 will result in a larger rejection region, making it "easier" to reject the null hypothesis. Thus, she chose answer choice *c*. However, Interviewee 169 did not state the correct rule. He claimed that larger significance levels correspond to *smaller* the rejection regions. Therefore, increasing the significance level from 0.05 to 0.10 will make it "easier" to *retain* the null hypothesis. Thus, he chose answer choice *a*.

Another "rule" quoted by some of the interviewees used to justify their choice involved α and Type I error. These interviewees relied on the fact that α gives the probability that the null hypothesis will be rejected, given that it is actually true, to justify their choice. Interviewee 172 used this definition in her reasoning for choosing answer choice *c*.

> **Excerpt: Interview 172**
> I: Ok. Um…oh ok well, um…I chose *c*…because…the way I understand the alpha value is the probability of…um…re, incorrectly rejecting the null hypothesis. So, that, *c* like actually, actually says that, so. Um…and because 0.1 is greater than 0.05 then it would be a greater chance of incorrectly rejecting the null hypothesis the, I guess that was just based on what I had learned. Like, a principle or, I guess, or a concept.

Interviewee 172 merely remembered that α is the probability of committing a Type I error and looked for the answer choice that supported that definition. Thus, she chose answer choice *c*.

Interviewee 192 also used the definition of Type I error in justifying her decision to change from answer choice *a* to answer choice *c*. However, throughout the discussion she became confused about the definition of Type I error. The following excerpt illustrates her confusion and use of the definition of Type I error.

> **Excerpt: Interview 192**
> I: Yeah, *a*, um… I think it goes back to the confidence interval thing? The way I see it, the bigger the…or the larger the alpha, the larger your confidence interval. I could be wrong but that's what I was thinking. Um…so, the larger your confidence interval the greater the chance of your null hypothesis falling in that interval and there's no way to know whether or not it's right or wrong since that is your acceptance interval.
>
> M: (Writing) Larger alpha means larger confidence interval?
>
> I: Um hum.

M: (Writing) And a greater confidence interval makes the null…there's a greater chance the null's gonna fall into it.

Here we see that Interviewee 192 used similar reasoning as that of Interviewee 169 in her initial choice of option *a*. Over the course of the discussion, however, Interviewee 192 changed her mind and used the fact that α is the probability of a Type I error to justify her decision to change her answer choice *c*.

Excerpt: Interview 192 (continued)
M: Ok, so how does that differ from *b*, *c*, and *d*?

I: Um…(reads *b* out loud). Um (reads out loud, somewhat inaudibly)…I think it…it goes into Type errors…like Type I error, Type II error, Type III, and Type IV error. And then type…ah… *a* is Type I and I know that's being…that's the worse one you can have…so.

M: Oh, *a* meaning, letter *a* or choice *a*.

I: Choice *a* is, um, Type I error. [M: Ok, and…] Do you really want me to go on to *b*, *c*, and *d*? (Laughs)

M: Yeah…(laughs)

I: Um…*b* would be…(reads out loud, somewhat inaudibly) I really don't know which one *b*, *c*, and *d* fall into…but I remember the teacher or the substitute talking that day, as going into the whole case thing. So like the worst… Type I would be like locking an innoc, innocent person up. That's how I think of it (inaudible). I really don't remember the rest.

discussion continues

M: But you like *a* better because?

I: Yeah. Because it falls into the… it…it's the Type I error. That's what struck me…like it's wrong but it's right. (Both laugh)

M: Ok. All right.

I: That they will…or *c*…that they will incorrectly reject the null hypothesis? Meaning that the…it was right and it….and you, and you didn't accept it. That's wrong. I think I choose that…I think I might have gotten *a* and *c* confused now that I'm thinking about it. Because if it was true…and you said it's not true. I think that's kind of worse…worse situation to have.

M: Than *a*?

I: Than *a*.

M: So you're kind of evaluating these based on what's a bad situation?

I: Yeah, like honest, honestly I can't really tell you the definition like I remember reading them. But, I don't remember them. Like what she said to me in the classroom struck me…like was more memorable to me then the um…than what I read in the book. So that's what I used to base it on.

M: Ok.

I: When doing my work, it's usually right…so (laughs).

M: Ok.

Interviewee 192 remembered having learned the definitions of Type I, Type II, Type III, and Type IV errors. When asked to explain why she did not chose *b*, *c*, or *d*, she was reminded of those definitions and tried to fit each of those answer choices to one of the error types. Ultimately, she chose option *c* because it was the worst error one could make. It seems that throughout the discussion, she lost sight of the question itself. However, the discussion illustrates the fact that she did not have a strong, resilient understanding of the concepts involved.

Analysis

With regard to the **Analysis of the Sample** and **Decision Rule** items, the data collected in the follow-up interviews confirmed the findings of the quantitative phase. The explanations offered by the interviewees indicated that they did not have strong understandings of sampling distributions or levels of significance. In fact, many of the interviewees did not even know what the term sampling distribution meant. Furthermore, the interviewees did not understand the role of sampling distributions and/or significance level in the overall logic and reasoning of statistical hypothesis testing. Often, the explanations offered by the interviewees were incorrect and, when their explanations weren't incorrect, they were merely statements of "rules" without reference to the concepts and reasoning that support those rules.

In fact, it was not clear that the interviewees understand that statistical hypothesis tests employ indirect logic to determine whether or not the collected sample is unusual under the null hypothesis. Hence, it is not clear that the interviewees understand the role of probability and, more specifically, of sampling distributions in determining whether the collected sample is unusual, conditioned on the null.

However, there are limitations to the analysis conducted on the two items representing these categories. These items appeared late in the assessment and were covered late in the interview. Often, additional probing questions were not asked about these items because (a) the hour was coming to an end and the interviewees were beginning to fatigue and/or (b) these questions had been asked in relation to other items. Therefore, before drawing broad conclusions about student understanding of the logic and reasoning of statistical hypothesis testing, it is necessary to consider student responses to the remaining items.

Recognizing Applicability and Conclusion Categories

Items from the **Recognizing Applicability** and **Conclusion** categories were written to assess whether introductory statistics students understand the value of statistical hypothesis testing as a method by which inferences about a population can be made based on information from a sample. In particular, these items assess whether introductory statistics students understand that the method offers a means by which an inference, or conclusion, can be made about the legitimacy of the null hypothesis as the true descriptor of a population, given sample information. In so doing, these items assess whether students understand the concepts and reasoning that support the development of a conclusion statement (reject or fail to reject the null hypothesis). In addition, the items in these categories assess whether, given the data and ultimate conclusion statement associated with a particular statistical hypothesis test, introductory statistics students can make valid claims about that information.

In the multiple-choice assessment, performance on the items representing the **Recognizing Applicability** and **Conclusion** categories was relatively weak. The quantitative results indicated that introductory statistics students understood statistical hypothesis tests to be a measure of the truth of the null or alternative hypotheses. Additionally, the results indicated that some students believe that statistical hypothesis tests prove one or the other hypotheses to be true and/or that some students think that sample statistics provide direct measures of the population parameters. Overall, the results indicated that introductory statistics students do not understand the value of the method nor do they understand what inferences (conclusions) are valid as a result of applying that method.

However, as was the case in the proceeding sections, the results of the quantitative phase tell only part of the story. The follow-up interviews provide a venue by which these issues may be explored further so that we may better understand student thinking about these components of statistical hypothesis testing. As a result, one item from the **Recognizing Applicability** category and two items from the **Conclusion** category were included in the follow-up interview.

Item number 3 was included in the follow-up interview. This item is classified in the **Recognizing Applicability** category and assesses student understanding of the value of statistical hypothesis testing as an inferential method used to make inferences about populations based on data from samples of those populations.

Figure 5.4

Item Number 3, Multiple-Choice Assessment

3. Which of the following statements is the ***best justification*** for using a statistical hypothesis test to answer the question: *Are female students as successful as male students in college mathematics*?

a. It allows you to talk about uncertainty associated with your decision.

b. It allows you to use only a sample of students to prove something for all students.

c. It allows you to calculate means and use statistical tables to answer the question.

d. It allows you to find and prove the answer using mathematical calculation.

Answer choice *a* is the correct option. Answer choices *b* and *d* should be eliminated because statistical hypothesis tests are not methods of proof. Answer choice *c* is not the best reason for using statistical hypothesis tests.

In the multiple-choice assessment, only 11.5% of the students chose the correct option; 37% chose answer choice *b*; 25% chose answer choice *c*; and 30% chose answer choice *d*. Of the eleven interviewees, 2 chose *a*, 4 chose *b*, 3 chose *c*, and 2 chose *d*. Over the course of the interview, one interviewee changed his answer from *b* to *d* and one from *c* to *a*. Ultimately, then, 3 people had chosen *a*, 3 people chose *b*, 2 people chose *c*, and 3 people chose *d*.

Interviewee Explanations: A Summary

The explanations offered by the interviewees for their choices on this item provided valuable information about their understanding both of hypothesis testing and of the language used in the item itself. Of the eleven interviewees, 7 did not choose *a* (the correct answer choice) because they were uncomfortable with and/or didn't understand the use of the word "uncertainty". This difficulty is illustrated in the following excerpts:

Excerpt: Interview 169
I: I didn't choose *a* because nowhere in the…in this…the question within the question does it ask about…ah…confidence intervals or how certain are you, type of thing….

M: Ok. In *a*, what about that made you think that that…cause you said that it, it doesn't…in the question they're not talking about confidence intervals…so, what prompted you in *a* to think that it should…that it was talking about confidence intervals?

I: Ah…when we started ended it…we started getting into, ah, 95%, 99% whatever confidence intervals and then some of the follow-up questions on those questions would be how certain….you know, comment on the strength of the evidence. And, the strength of the evidence, it's, you know, the stronger…the more certain you are there, you, strong, the stronger the evidence for your conjecture, whatever it is.

Excerpt: Interview 192
I: The major differences…um…well…*a* is, talks about uncertainty…and, honestly, I don't like that word, especially when it comes to math. So I kind of stay away from it.

M: Why don't you like that word?

I: Because math is usually black or white and uncertainty is that gray in there. It's like ah…it's not…I'm pretty sure in…with anything regarding math, uncertainty must not be right. Um…*b* allows you to use a sample of students…um…we've been doing that all year…kind of…getting…using small populations to assume things about the larger populations.

Excerpt: Interview 77
I: Um…I don't…I think I didn't pick *a*, because I wasn't sure what it meant by like I guess, the uncert, uncertainty associated with your decision. Um…I mean I guess that's talking about…like….there's a possibility that female students aren't as successful as male students. Um...but that was just…I didn't really know what that was talking about.

Excerpt: Interview 191
I: *A*? Allows you to talk about uncertainty associated with – um, well, I'm a little ah… unclear on the, talk on, on the phrase, "Talk about uncertainty." I don't know exactly, ah…I mean I guess that might have something to do with maybe standard deviations or, um, how accurate the data is, but, ah,… I don't know through hypothesis testing you can, um….where that information is really critical. Yeah, 'cause when I think of uncertainty, I think of um, you know, correlations or um how the data is spread apart,

> and, ah, I don't think of that…I don't remember that being, um, too important or even used really in the hypothesis testing.
>
> *discussion continues*
>
> I: Well, I think you would just calculate standard deviation or, ah, correlations with that, um…'cause again, when I read, "Uncertainty," that's what I think of, data or, um,…values that'll, um, indicate, ah, you know, how the data is spread apart, um…any correlations whether, you know, female students are as successful as male students. Um…I just don't remember that being a reason, um…or that being answered when we did hypothesis testing.

The responses indicate that these interviewees did not associate the word "uncertainty" with statistical hypothesis testing. Interviewee 169 only associated the word "uncertainty" with problems that involved confidence intervals. Interviewee 77 did not understand the use of the word in this context. Interviewee 191 only associated uncertainty with spread of data, not with using hypothesis tests to make an inference. Interviewee192 thought that statistics is mathematics and since "uncertainty" is not a concept associated with mathematics, it must not be a concept associated with statistics, either. Her comments are interesting, but problematic because statistics is <u>not</u> mathematics. In fact, it is the ability of statistics to deal with variability, chance, and uncertainty that distinguishes it from mathematics. As we will see, this issue arises with other interviewees at other times.

Interviewee 29 chose the correct answer, option *a*. However, she did not choose that option for the correct reason. Her explanation indicates some confusion over the use of the word "uncertainty".

> **Excerpt: Interview 29**
> I: Um… I think I chose this one because…um…because I feel like we used hypothesis tests…I guess it was try to see if…like…um…say you think…or you guess, like, you're not sure if something is…is like, I don't…like the one group has like the same characteristic as the other. So, I guess you use hypothesis testing to see if there is a difference. So, that's why I picked *a* because you're not sure about something so you're applying, ah, these statistics to see if there's actually a difference and…using math.

> M: So you use hypothesis testing, it's hard to say all that [I: Yeah (laughs)]. You use hypothesis testing when you're unsure about something?
>
> I: Right, because then maybe like the numerical answers can support like…like, you know, your original thought, or guess or something.

Though Interviewee 29 chose the correct answer, her reason for doing so did not indicate an understanding of the role of hypothesis testing in drawing an inference about a population. Rather, she chose this option because she associated the word "uncertainty" with the uncertainty a researcher has *prior* to conducting the study: uncertainty about the answer to her question. It is this initial uncertainty that prompted Interviewee 29 to choose answer choice *a*, not the uncertainty associated with the ultimate conclusion made at the end of the test.

Many of the interviewees chose either answer choice *b* or answer choice *d*. Unfortunately both *b* and *d* are incorrect options because statistical hypothesis tests are not methods of proof. They can not prove the null hypothesis nor can they prove the alternative hypothesis. In the quantitative phase, if a student chose one of *b* or *d*, it was inferred that the student believed statistical hypothesis testing to be a method of proof and, furthermore, that the student believed the best justification for conducting a statistical hypothesis test is that it provides a proof of the null (or alternative) hypothesis. However, in the explanations offered by the interviewees, it is not clear that this inference about student understanding is correct.

The interviewees cited a wide variety of reasons for choosing either *b* or *d*. Consider the following excerpt from Interviewee 77:

> **Excerpt: Interview 77**
> I: Ok…um…I think I chose *b* from what we like learned about from…if you have like a large population sample…or, not sample, a large population and you want to test a certain claim…then you can take a smaller sample from those populations, um…that

> and…do a statistic…or statistical test…um…to…prove or disprove that claim. And, that's justified because if the population's large enough, then you can, um, assume that it's like normal…or normally distributed. That's why you can use the sample. Um…

As Interviewee 77 continues, she explains that *c* and *d* are very similar. Since they were both so similar, and since she could only pick one, she went with the option that was most familiar, option *b*. She did not *necessarily* choose option *b* because she thought statistical hypothesis tests prove or disprove a given hypothesis. Instead, she used test taking skills and chose the option that was most familiar to her.

Interviewee 192 gave a similar reason for choosing answer choice *b*. She chose the option that was familiar to her.

> **Excerpt: Interview 192**
> I: Ok. (Reads number three out loud). Um…(reads out loud, somewhat inaudibly)…hmm…honestly I don't know why I chose that one.
>
> M: Which did you choose? *B*?
>
> I: *B*, yeah. I might have chose it because it's the only thing that looked familiar to me. Cause even looking at now, I (inaudible) (laughs).

Because option *a* was eliminated (due to the word "uncertainty" – see excerpt included above) and because *c* and *d* were not familiar to her, Interviewee 192 chose option *b*. She remembered that when she solved statistical hypothesis testing problems in class, she used sample data to draw a conclusion about the population.

Interviewee 172, gave quite a different reason for choosing answer choice *b*. Her reasoning is illustrated in the following excerpt.

> **Excerpt: Interview 172**
> I: Ok. Ok. I chose *b* for this one. And…ok…I guess…I'm trying to think…um…I guess I, you know, knew from this that you could do, you would have to take a sample from both female and male students to be able to apply to hypothesis testing. And…I'm not sure why I chose this and not the other ones, cause I guess the other ones could…yeah, I…they also apply but…um…and I know that there would be uncertainty in…this test because it would only be a sample of the students and not the…general population.

> M: And so *c* and *d*? You think they're ok or…because the question asks what's the best justification [I: Right]. So, it could be that three of them are reasonable but [I: Um hum] only one is the best or it could be that two of them are reasonable or…there's only one that's actually could even be considered. So what [I: Right], where do you, what camp do you see yourself falling into?
>
> I: Well…I mean, I guess that I could, *c* is reasonable because if you took a sample of the students then you can figure out the…the mean and all the other, ah, like standard deviation, things like that from your sample. And, I guess, then apply it to the…all female students or all male students. Um…and then…I don't know, I guess after, I mean, you would have to do all of that calculating and then afterwards, you know, be able to come to a conclusion…that, I guess. I guess I kind of thought about it in like a sequence of events that…you'd have to chose your sample first and then…do the calculations and then actually prove the answer.

Interviewee 172 thought that all four answer choices were reasonable. However, she could choose only one. Subsequently, she chose the statement that described the first step in statistical hypothesis testing. Once again, we see evidence that students who chose option *b* did not *necessarily* do so because they thought statistical hypothesis tests were proofs.

This focus on aspects other than the word "prove" was also evident in the explanations offered by the interviewees who chose *d*. Consider the following excerpt from Interviewee 122:

> **Excerpt: Interview 122**
> I: Three? Well, I chose *d* from all the answers because, um, *d* is – it allows you to find and prove the answer using mathematical calculation. Well, because if you gather all the data from each female and, um, male, you could find the mean, median, and the mode regarding to the – and also make those, um, uh, diagrams like the histograms go box-and-plot – box-and-whiskers plot and then be able to make you have a decision also. Yeah, because, um, I was like, regarding to this, it kind of goes with letter *d* because regarding mathematical calculation, that deals with mostly a whole range of math. Um, like I said, the mean, you could find the standard deviation, you could find the mode, you can find everything thing just regarding to this mathematical calculation. So, all of, like, *c* falls into *d*. And the other two, on *a* it allows you to talk about uncertainty associated with your decision, well this falls also with *d* because you're trying to make a decision, but you're proving your answer. And *b*, um, it allows you to use only a sample of students to prove something for all students. Well, that also kind of falls into *d* because it does – you need to find a particular, um, sample to actually make mathematical calculations. That's my opinion.

Here we see that, like Interviewee 172, Interviewee 122 thought all four options were reasonable justifications for using statistical hypothesis tests. In order to choose only one option, Interviewee 122 relied on her overall understanding of statistical hypothesis testing. She understood statistical hypothesis testing to be a process that relies on a series of calculations to arrive at a conclusion. Each of the options discusses *aspects* of statistical hypothesis testing. But, for Interviewee 122, the statement in option *d* sums up the test and its value. Though Interviewee 122 used the word "prove", it was not central in her reasoning. Rather, she depended on her understanding of the method and its uses to choose answer choice *d*.

Interviewee 81 also chose answer choice *d*. In fact, he initially chose *b* and, over the course of the interview, he changed his answer from *b* to *d*. His reasoning is highlighted in the following excerpt from that interview:

> **Excerpt: Interview 81**
> I: Ok, yeah I chose *b*…um...it allows you to use only a sample of students to prove something for all students. Um…I think, hold up…oh. I might have misread the question. I'm not sure though. Ok…um…I would probably…I'd say *b* or…*d*. Or, I think I chose *b* because you're using like a sample of students to prove something for all students….cause ah…because…um…statistical hyp…because statistical hypothesis testing, you can't test everybody. You can only test, like…so you can't test everybody in the world, or everybody in college. Yeah, you can only test, like a certain number. So, I think that's why I went with *b*.
>
> M: Ok. And…you now like *d*…why?
>
> I: Um…well because it allows you to find and prove your answer using mathematical calculation. And, like…cause, like in any math you have to use calculations, right? And, um since you're sort of testing, you know female students and male students, you would have to have calculate which is…if they are as successful. So you would need to use mathematical calculation to do the statistical hypothesis for this question.
>
> *discussion continues*
>
> I: Whether that's…*a* is the answer or not…I don't know…but I don't like it personally. Um…I would probably go with *d* if I was to change my answer…allows you to…ah…find and prove the answer using mathematical calculations. That's why you, you kind of use statistics. You use math, mathematical calculations to prove something

> wrong or right. So, I kind of like *d*…um…if I were to change my answer, which I probably would if I were to retake the test, I guess. So…

Initially, like Interviewees 77 and 192, Interviewee 81 chose answer choice *b* because it represented something that was familiar to him. However, he changed his answer to choice *d* because, for him, it described the goal of statistical hypothesis testing. Again the word "prove" was not a central idea in his choice of an answer.

Interviewee 15 also chose answer choice *d*. However, her reasoning for doing so is interesting in that she explicitly did not pick option *b* because she did not think that statistical hypothesis tests "prove" a given hypothesis but then chose *d*, which also uses the word "prove".

> **Excerpt: Interview 15**
> I: I guess I chose *d* because it's like really straight forward. Like…if you're doing mathematical calculations, like…that's something concrete. Um…cause like if you're just, if you're talking about uncertainty, like hypothesis testing is (inaudible) give you a lot of stuff about uncertainty cause like uncertainty with that would be stuff like um …you don't know all the variables for female students versus males students. And that's not really math. And, then, like you can never really prove something for all students, like you can suggest something that would apply to all students. But, you can't really prove it unless you sit there and talk to every single student (inaudible).
>
> M: Hold on, let me get some of this down here. I want to just double check because I'm afraid that that won't pick up everything. Ok, so…uncertainty for *a* is not mathematical?
>
> I: Yeah, not…yeah, not really most of the time, I guess.
>
> M: So *a* is just not concrete enough?
>
> I: Yeah.
>
> M: Ok, and then what were you saying for *b*?
>
> I: Um, I guess it's just unlikely that you like could really prove something for all students. Because like the hy…like different things let you suggest things that are probably true for all students. But like…um... just like the confidence intervals, that's why you use them because it's not like absolute. You're just suggesting and throwing it out there.
>
> M: Ok. And, *c*?

> I: *C*? I mean I guess like…I think calculating means and using statistical tables is kind of part of a hypothesis test but you can do that without using a hypothesis test. Like you don't need to use it for that.
>
> M: Ok. So do you see any similarities between the choices, or not?
>
> I: Yeah, um, I guess like they're all like really close to being an answer…like calculating mean that's part, that's like because you need the mean to be able to do the hypothesis test. But like I think you could also just (inaudible) I mean not do the hypothesis test too. And, you …you are using like a sample but like… then again like you are not really proving anything concretely unless sit there and go through every single student. Then you'd be able to say that you could prove it.

This excerpt highlights the fact that, though she picked answer choice *d*, interviewee 15 does not believe statistical hypothesis testing to be a method of proof. She focused on the word "prove" in her decision to eliminate answer choice *b* but the word "prove" did not have the same impact in her consideration of answer choice *d*. Instead, she chose answer choice *d* because it was "concrete". Option *c* presented a justification that might be applied to other concepts. Option *a* presented a justification based on the concept of "uncertainty", a concept with which she was uncomfortable in this context. Note that, once again, we have encountered an individual who associates mathematics with statistics. Ultimately, Interviewee 15 chooses option *d* as it is the statement that represents the more concrete, mathematical ideas she associated with statistical hypothesis testing.

Of the 5 interviewees who did not choose options *b* or *d*, only 2 of them eliminated *b* and *d* explicitly because of the word "prove". These two interviewees explained that nothing can be proved using statistical hypothesis testing unless all members of the population were included in the analysis. Two of the remaining three eliminated answer choice *b* because the statement did not clarify the *type* of sample and eliminated choice *d* because mathematical calculations are used in other statistical methods, not just statistical hypothesis testing.

In summary, though the word “prove” was explicitly included in two of the distractors to assess whether students believed statistical hypothesis testing to be a proof, this was not the deciding factor used by the interviewees in choosing their answers. The issue of proof was not central in the justifications interviewees gave for choosing answer choices *b* or *d*, nor was it central in the justifications interviewees gave for eliminating answer choices *b* and *d*. The fact that the interviewees did not focus on word “prove” in their reasoning for item number 3 challenges the conclusion posited by the results in the quantitative phase. This issue will be explored further in the analysis of student explanations for choosing (and not choosing) various answer choices on items five and six from the **Conclusion** category.

Item number five assesses whether students are able to interpret the conclusion to a statistical hypothesis test in which a significant difference is found.

Figure 5.5

Item Number 5, Multiple-Choice Assessment

5. In 1950 the mean IQ of undergraduates at a university was 110. To test the hypothesis that students today are smarter, a study of 500 current students found a mean IQ of 120. The difference between the two means is significant at the 0.05 level. ($\alpha = 0.05$)

Which of the following statements is necessarily true?

a. Undergraduates at the university today are smarter than those in 1950.

b. The claim that undergraduates today are not smarter than those in 1950 is true with a probability less than 0.05.

c. The claim that undergraduates today are smarter than those in 1950 has been established with 95% certainty.

d. If undergraduates today are no smarter than those in 1950, the probability of the observed mean IQ is less than 0.05.

Answer choice *d* is correct. Answer choice *a* was written to assess whether introductory statistics students believed that statistical hypothesis tests constituted proofs of hypotheses. Answer choice *c* was written to assess whether introductory students believed that, when a result is statistically significant, the significance level can be used to find the probability that the alternative hypothesis is true. Answer choice *b* is a variation on *c* but states the significance level directly, as a probability.

Overall performance in the quantitative phase on this item was low. Only 6.7% of the participants chose the correct answer choice, *d*; 87.5% chose option *c*; 3.8% chose option *b*; and 1.9% chose option *a*. Because very few chose *a*, it was inferred (contrary to what was suggested by performance on item number 3) that introductory statistics students do not believe statistical hypothesis tests to be methods of proof. With the overwhelming majority of students choosing option *c*, however, it was concluded that introductory statistics students believe that statistical hypothesis tests provide a measure of the "truth" of a given hypothesis.

Of the eleven interviewees, none of them chose *a* or *b*, 9 chose *c*, and 2 chose *d*. Over the course of the interview, one person changed his answer from *c* to *d* and one changed from *d* to *c*. Ultimately, then, the frequencies with which answer choices were picked remained the same as they were prior to the interviews. The follow-up interviews provided more insight into student thinking about the meaning of the conclusions reached by statistical hypothesis tests.

Interviewee Explanations: A Summary

Each of the interviewees did not choose answer choice *a* because it was too broad a statement. However, there was variance in the way that the interviewees thought answer choice *a* was too broad. Two interviewees noted that *a* was not correct because it made a broad claim about a population based solely on information about a sample.

> **Excerpt: Interview 15**
> M: So what do you think…ok, so why didn't you choose *a*?
>
> I: Um…cause you're looking at just a sample of students. So like you can't like really come out and say like these students are definitely smarter than those students cause you didn't really look at all the students. So you can't really say anything concrete cause there's the possibility that the other 500 students at the university that you didn't test are geniuses and …or like are really dumb or something… so then you won't know cause you never tested those other students so you can't really say anything like with concrete certainty.
>
> **Excerpt: Interview 77**
> I: Um…letter *a*. Since the study only tested 500 current students, I mean there could be a chance that those 500 students were, I guess, smarter than the average…um, average student. So, I felt that *a* was too general to say…or to assume that was true.

These statements indicate that interviewees 15 and 77 understand that, because a sample is being used for the analysis, the test does not necessarily constitute a proof for the entire population. Thus, they did not pick answer choice *a*, as it made a broad, general statement about a population based solely on information from a sample.

It was not so clear from the other responses that everyone had this understanding. Eight of the eleven interviewees claimed that they did not pick *a* because the statement was incomplete. It did not provide information on the level of certainty associated with the test. Some examples of this line of reasoning are included in the excerpts below.

> **Excerpt: Interview 132**
> I: Ok. (reads out loud, somewhat inaudible) Um…well *a* says like the same thing…I chose *c*. But…ah…but *a* says the same thing except it doesn't have…an…a certainty level. So that means it's missing out information, cause here we can only know to a 95% uncertainty. Um…so that's why I chose *c*.
>
> **Excerpt: Interview 191**
> I: Hum. Ok. All right, well *a*…Ok, reading this, *a* is not necessarily true because you can test at different, um…ah, levels. I guess alpha, um,…I think that's the critical value or…I don't remember. Basically, it, it…the question tells you what alpha level it's being tested at, and that can change, um, so because that can change depending on what alpha you test at, um, this isn't necessarily true, where undergraduates are smarter than those in 1950. Um,…(reads options somewhat inaudibly). Ok (Laughs). I think I chose

> *c*, undergraduates today are smarter than those in 1950, has been established with 95% certainty. Ah…yeah, I chose that because, um, the alpha value, ah, corresponds with this level of certainty.
>
> **Excerpt: Interview 29**
> I: I guess, I chose…I didn't choose *a*…I chose *c* because it said with 95% certainty…and that was just, I guess, being more specific. Um...I don't have a really good answer, it's just basically because it's just more specific to the problem.
>
> M: So you do think *a* is true? It's just that *c* was more specific? Or, you don't think *a* is true?
>
> I: I think, *a* is true.
>
> **Excerpt: Interview 122**
> I: Okay. Undergraduates at the university today are smarter than those in 1950. They're just saying that, um, they're just stating the fact that the students of today is just – has a higher IQ than – than the 1950. It doesn't tell anything else about the – the level of significance. That's why I, um, kind of crossed that out.
>
> M: Okay.
>
> I: Because every time when your trying to make a statement you need to know what your alpha is, your degree of sig – uh, certainty is.
>
> M: Why? Why do you need to know that?
>
> I: Because then whoever's like reading it knows what kind of, um, populate – like, um, the kind of like the region or area where it's, um, it should lie between, their scores.
>
> **Excerpt: Interview 81**
> I: And, I thought it was *a* or *c*. And, I thought it was at the 95% certainty level. So, that's why I chose *c*, cause other, other than that, *a* and *c* are pretty much the same.

For these interviewees, it was important that the final statement of the conclusion to a statistical hypothesis test be accompanied by some indication of the degree of certainty associated with the test. Therefore, they did not choose option *a*. As we will see, these interviewees believe the degree of certainty is related to the level of significance used in the test.

Of those who chose option *c*, the reasons given for eliminating *b* and *d* were varied. Some of the interviewees claimed that *b* and *d* were difficult to understand.

> **Excerpt: Interview 191**

I: …Ah…and ah…and yeah, *b* and *d* ah…a lot of things going on in those answers [M: (laughs)]…and so, yeah, I'm having trouble understanding them.

Excerpt: Interview 169
I: Yes. Um…*c* looks…to me the most right between *a*, *b*, and *c*. *D* doesn't make a lot of sense to me…so. Um…I chose *c* over *b* because…you're not proving, when you're hypothesis testing, to my knowledge, you're not proving that…something's true. You're claiming that they are with a 90, with a certain amount of, ah…strength with your evidence. And, in this case the strength would be 95% certainty.

Excerpt: Interview 77
I: So *b* is saying that there's a less than…0.05…I guess percent chance that the claim is not true, since the hypothesis was testing undergraduates today are smarter. That one's kind of confusing to me. Same with *d*. I'm trying to…

When asked to try to express the meaning of option *b* their own words, interviewees 191 and 169 were eventually able to see that *b* and *c* were actually the same claim, stated differently. Interviewees 29 and 15 also made this realization. However, none of them chose option *b*. There was a general feeling that *c* made the statement more clearly. Therefore, the interviewees chose that option. The following examples illustrate this reasoning.

Excerpt: Interview 15
I: Um because we never really I guess like associated the alpha like with a certain probability and like I guess it could be true because like they found that they were smarter with a lot of certainty so claiming that they're not smarter with a really low probability seems like it would be similar. But, like *c* was more straightforward and made more sense.

discussion continues

I: Yeah, like…cause it's like…it's like too many negative words…like no smarter and like less than. So it's like harder to sort it out, I guess. You gotta like really sit there and like stare at it to sort it out and make it mean something. Like, *b*…it said like not smarter and less than. So it seems like you have to flip it around before you can figure out what it's really saying.

M: So once you do figure it out what *b* and *d* are saying, you still don't think they're true?

I: Um, I just feel more confident with *c*.

Excerpt: Interview 169

I: The more I actually read *b* and *c* the more *b* looks like just the opposite of *c*. Um…I must have misread…I must have sped read through…not taken the whole time, but I still think that...they just seem like opposites to me.

M: But, you still like *c* better?

I: Yes, I don't think I've seen something phrased in the way that *b* is phrased. It's like a…a negative answer…a negative answer in that it's asking for the opposite of something and then, again, with a less than…I'm not really…it just seems like it's asking for two negative versus two positive things in *c*.

Though Interviewees 15, 29, 169, and 191 were able to understand the claim made in answer choice *b*, the statement in option *d* remained somewhat mysterious to them. Thus, they did not choose that option and, instead, chose answer choice *c*.

Other interviewees appealed to the sample data to justify the elimination of choices *b* and *d*. Because the sample data supported the claim that the current students are smarter than those of 1950, Interviewees 81, 122, and 132 eliminated options *b* and *d*. This reasoning is illustrated in the following excerpts.

Excerpt: Interview 81

M: Ok, so what are…what is *b* saying, in your own words?

I: The claim that…oh, basically it says that people today are not smarter than the people back then. And that doesn't make any sense because it says the IQ is higher today than it was back then.

M: And, what is *d* saying?

I: They say they are no smart…and that means like equally as smart and that was proven wrong because the IQ is higher than it was back then.

Excerpt: Interview 122

I: Um-hum. Okay, *b*, the claim that undergraduates who – undergraduates today are not smarter than those in 1950 is true above probability 0.05, which I disagree because, um, they actually, the undergraduates today is greater than, um, 1950.

discussion continues

M: And *d* you don't like now because –

I: Um, uh, if undergraduates today are no smarter than those in 1950, the probability of observing IQs less than 0.05, well for that one, is, um, it's kind of just, like,

> it – the – the undergraduate today are no smarter than those – well, they kind of are smarter than them because regarding to the, um, the answer, like the mean for 120 versus 110. So you can tell there is a graduate – there's some increase.
>
> **Excerpt: Interview 132**
> I: *C* would mean that they're accepting the alternative hypothesis. They're saying that present…ah…graduates nowadays…would…are smarter than those from the 1950s. And, that is true to a 95% certainty so that means there is 5% …um…there's 5%...ah…area that you could doubt the results. For the most part they're certain that it's right.
>
> M: So how is that different from *b*? Or is it?
>
> I: Um…well in *b* they're accepting the null hypothesis and in this one, they're accepting the alternate hypothesis.
>
> *discussion continues*
>
> I: Because…ah…the data like ah…the mean from now is higher than the …than the mean from the 1950s, so. I didn't go through the math [M: Ok] but it looks like the data is leaning towards… For some reason I thought we couldn't do math when we took the…
>
> M: Well, that's true…that's right. You didn't have to do any calculations. It was all supposed to be…Ok, so because the data supported the null…
>
> I: Ah…supported the alternate.
>
> M: Or, the alternative
>
> I: That's why I chose…um…*c*, which is alternate.
>
> M: Ok. I'm just trying to work through what you…how you were thinking about this. So you saw…you saw *b* and *d* as being about the null?
>
> I: Um hum

In their explanations we see that Interviewees 81, 122, and 132 focused on the claim offered in the beginning of each statement. They did not focus on the probability associated with each statement. They were mostly concerned with whether or not the data supported the initial claim of the statement. Hence, they chose answer choice *c* because, in contrast to *b* and *d*, the sample data supported the claim being made in the beginning of that statement.

Overall, for those who chose answer choice *c*, there was strong discomfort with option *d*. This result is interesting in that answer choice *d* is the correct option. Two interviewees, however, finished the interview having chosen option *d*. Unfortunately, their explanation for doing so was not correct. Consider the following reasoning offered by Interviewee 172:

Excerpt: Interview 172
I: Ok. Um…I guess I chose this one…because I was kind of…thinking of…what the…significance level actually meant…and…I'm trying to think back now, but…I knew that the…the *p*-value is based on that and the…*p*-value is where you could ….the smallest number you could pick and still reject the null hypothesis, I think. And…um…well…I don't know. I think I didn't pick *a* I don't think because…it only says that it's significant at the…this level. So, I mean, I didn't know if that meant that…it really wasn't…that I guess the…it wasn't really any different than, like whatever your test statistic ended up to be, it didn't fall in the rejection region or it did. So, you couldn't tell whether…ah…today's students were smarter or not. And, then….I think I was probably deciding between *b* and *c*. And…I don't know just…I, I mean *b* and *d*, I'm sorry. I don't think *c* sounded right to me. I just…I…for some reason I didn't think it had anything to do with, the amount of, the percent certainty, I guess. But, like that confidence associated with it, but…

M: You didn't think it had anything to do with the confidence?

I: Yeah, well it's like it's been established with this 95% certainty. I just, I don't know, that sounded like it didn't…um…have to with it. I don't know, it just…it didn't sound right to me. And then…*b* and *d* do kind of try, I mean, I guess, the way I was thinking is that it related somehow…the…um…to, I guess, it being accepted or rejected. Um…based on the level…so. I'm not sure why I decided on *d*…but…that's at least how I got down to those two.

discussion continues

I: Um…well…the…0.05 in this case since it's…the alpha value it's…that is the probability of making a Type I error, which I guess, which is when you reject the null hypothesis although it's true. So, again, you're just …just like in the, when…if you're just doing confidence intervals or any type of approximation, I guess to the normal distribution. And, alpha is used…ah…it's all about, it's still a probability here. So, that's, I guess what led me to *b* and *d* that it was referencing this as a probability. Um…and…I guess…um…in this case it would have to…whatever the…um…*z*-value turns out to be…it would be some…like the *z* is greater than or equal to *z* of 0.05 so…I guess…ah…you…find this and then based on the test statistic…um…it…I guess…if it's significant at this level then that means either it's accepted or rejected but…I don't know because…I wasn't sure whether it was or not. I think

discussion continues

> M: So…and why did you like that one better than *b*?
>
> I: Um…I don't know, I guess because this one did provide, like, at least the *if* undergraduates today are smarter than those in 1950 which makes you think either it's um…I don't know, at least that gives you some indication that, in this case it, at least in this selection that it was…um…retained. Whereas the other ones, I mean…*b* just kind of…I don't know (reads *b* somewhat inaudibly). I don't know, I guess this *d* just gave me some indication of accept or reject so it was easier for me to relate the…alpha equals 0.05 to the testing.

The reasoning that Interviewee 172 used to arrive at option *d* is not indicative of deep understanding of the logic and reasoning of statistical hypothesis testing. She did not associate the word "certainty" with statistical hypothesis testing (as the interview progressed, it became clear that she associated the word "certainty" with confidence intervals, not hypothesis testing). Thus, she did not choose option *c*. Though the item stated that a significant difference was found, she wasn't sure whether the null hypothesis was rejected or accepted. She remembered that alpha is the probability of making a Type I error. Both *b* and *d* included the words "with a probability of 0.05". Therefore, her attention was focused on these two options. However the statement in option *d* included the word "if" while the statement in option *b* did not. She, therefore, had a sense that the hypothesis might be rejected or retained in answer choice *d.* This seemed familiar to her so she chose it. Unfortunately, she did not choose it for the correct reason.

One other interviewee chose option *d*. As was the case for Interviewee 172, the reasoning used by Interviewee 112 for choosing option *d* was not correct. Initially, Interviewee 112 chose option *c*. But, as he reread the item he realized this was not the correct choice and changed his answer to *d*. His reason for doing so is illustrated in the following excerpt.

> **Excerpt: Interview 112**
> I: (reads out loud, somewhat inaudible) I put the claim that un…I put *c* the claim that undergraduates today are smarter than in 1950 has been established with 95% certainty. It…hmm…for this I had 90% certainty (laughs and reads out loud, somewhat

> inaudible) Well, I guess, you can't put *a* because there is like a percent error margin. (reads out loud, somewhat inaudible) I guess, I'll go back to *b*. (reads out loud, somewhat inaudible) Actually, *d* might actually be right. Cause, it probably does need alpha over two there.
>
> M: Why?
>
> I: I'm not really sure why but it kind of strikes me is that now though…yeah, because 0.05 you probably need the 0.05 at the beginning and the 0.05 at the end also. You need like .9 in the middle. So it's 90% certainty, not 95%.
>
> *discussion continues*
>
> M: So *a* you aren't concerned with. Right? You pretty much said no… ok. So what are *b*, *c*, and *d* actually saying? Can you say those in your own words, what they are telling you?
>
> I: (reads out loud, somewhat inaudible) Well…I guess it's saying that they're not…there's a 0.05%...there's a 0.05 probability that undergraduates today are not smarter…and that's not…that's not real…that's not necessarily true because they could also be equally as smart. Cause it's just the error margin that they're not smarter so it's leaving out the equal to part. And, then, *c*? It says, 95% certainty but with 0.05, it should be 90% certainty. And with *d*, if undergraduates today are no smarter than those in 1950, the probability …are no smarter…the probability of the observed IQ is less than 0.05. Hmm…(reads out loud, somewhat inaudible)…hmm…like *d* can almost be the same thing I said for *b*. I guess…(reads out loud, somewhat inaudible)…I guess cause it doesn't say is true like *b* says … that's probably why *d* seems more a better choice.
>
> *discussion continues*
>
> I: Um… think I'm gonna go with *d* now.
>
> M: And mainly because…?
>
> I: I think it's cause it doesn't have the is true part that *b* has. Because that kind of implies that there has to be less than. Whereas it seemed like without the is true, it leaves open the possibility of equal to in *d*.

Interviewee 112 was confused about whether it should be "alpha over two" or not. Though he later stated the hypotheses as if the scenario required a one-tail test, Interviewee 112 maintained that it "probably does need alpha over two there". He reasoned that, if alpha equals 0.05, the degree of certainty should be 90%, not 95%. Thus, he decided that option *c* is not correct. In considering option *b*, he was bothered by the fact that the statement merely said "probability that undergraduates today are not smarter". It did not include the "equals" in it. He might have been thinking that the null hypothesis should allow for the samples to be equal and therefore, did not believe this was a correct statement of the null hypothesis. He reasoned that, because the statement in answer choice *d* did not include the phrase "is true", it allowed for the possibility of equality. Therefore, he chose that option. Though Interviewee 112 chose the correct response, we see that his reasoning was not due to a deep understanding of the concepts assessed by the item.

Overall, this item was difficult for the interviewees to answer. Most of the interviewees did not choose the correct answer and, as evidenced by the data, it is questionable whether those who did answer correctly did so for the right reason. It is, though, an item that really addresses the heart of the logic and reasoning of statistical hypothesis testing. Because performance on this item was low, more time was spent discussing this item than was for others in the interview. In particular, probing questions were asked to determine the understandings students had of the concepts, logic, and reasoning that are drawn upon to reach a conclusion. Through these probing questions, an effort was made to gain insight into student understanding of how the numbers given for significance levels and the use of sample data connects to the overall logic of statistical hypothesis testing.

When probed to determine their understanding of how a conclusion is reached, the interviewees frequently referred to rejection regions, confidence intervals, and statistical tables and calculations. Many of the interviewees gave "rules" associated with certainty, rejection regions, and confidence intervals as illustrated in the excerpts below.

Excerpt: Interview 132
M: Like, how does this all connect? Where do these numbers come from…and…I mean…yeah, no not the math

I: The percentage?

M: Yeah.

I: Well cause whenever you do hypothesis testing you need to know to a level of certainty. Like you either you know 80% certainty or 99% certainty. So, like if you do like …ah…um…like a confidence interval. Like the longer your…the bigger your certainty, that the wider your interval will be. So the same thing…you would…um…you're…what is it? The area where you could still reject your null hypothesis becomes smaller and smaller with the more with certainty you have. Right you always have a level of uncertainty with the data. Like, to form your conclusions.

Excerpt: Interview 77

I: Right. Um…with the 95% um…like when we learned about confidence intervals…um…when you find a confidence interval….you take, if there's an alpha level given, then you take that from 100…um…to find, like, how confident that interval from…or how confident…you are…or how confident…trying to word this. Ok, so when you find the confidence interval you're, you're finding the interval for…um…the set of numbers in which the mean is contained. And…with…like say you find a 95% confidence interval and you find that the mean is contained in that interval of numbers then you're saying that you're 95% sure that the mean is within that confidence interval. So it's pretty much, um, much like the cert…it's like um…assuring a certainty in which like that number is in there…um…So when I think about this where it says with a 95% certainty, I'm thinking that…that study is 95% sure that the claim is true because the mean that they found was within the 95% confidence interval.

Excerpt: Interview 29

I: Um…I guess I think about it in terms of probability. So if there's like a higher percent…of…certainty that means there is a high probability of like that data like…or that characteristic occurring within…or, you know…like, it happening in the middle range, I guess. Um…so when there's like 0.05 or, you know, 0.1 or whatever…that's just basically like narrowing…or like, yeah narrowing the interval of …of that range of data…occurring. So (laughs)…so I guess it's just like a higher…like if it's greater than 0.05 or whatever then it's just…um…it's…I guess it's narrowing the interval? So, it just…makes it…less…ah (laughs)…um…sorry I'm confusing myself (laughs).

Excerpt: Interview 191

I: Um, I think, I think that has to do with the level of certainty. Um,…ok, so if you're using ah, a distribution, the normal distribution, the alpha, um, is sort of, I think it was the critical value. I think alpha was the critical value. Um, so, um, if you have your bell curve, um, your alpha would establish, um, how confident you were, um, that your, ah, answer was true, so depending on what value you get on this, it's a *z* distribution, I think, and you would get a value, um, that would either be – you, or sorry – so you would establish a value, um, among the first steps. You would establish a value that's a rejection, um, value, um, and that would be anywhere on this distribution, and then you would calculate, um, a value, not – so if that's your rejection region there, you'd calculate a value, um – I don't remember exactly how – that would either lie – that would lie somewhere in relation to this, um, so in this case, you're trying to test if it's greater than – so if your value that you calculated fell here at this line [referring to a graph of a normal curve], then it would be greater than the rejection region, but I think in that case you

> would accept the alternative. Um, now the alpha value, um, I think will tell you how certain you can be of your answer being correct, um. I remember alpha values are at the end of the distribution, um, and, ah, the smaller the alpha value, the more confident you can be that your answer's correct, um, because as the alpha value increases, then that means, like, if you have a smaller alpha value here, then that means you can be pretty certain of information for data that's represented within that region. As the alpha value grows, this region gets smaller, and I think that means that this information becomes – you're less confident that it's correct.

These excerpts are representative of the reasoning that most of the interviewees used to connect the notion of significance level to the decision to reject (or fail to reject) the null hypothesis. We see that generally, the interviewees recited a set of rules used to determine the conclusion to a statistical hypothesis test. These rules referenced technical terms such as rejection regions, *z*-tables, *t*-tables, and normal curves.

When probed to explain the terms they were using, the interviewees struggled. Again, it seemed they knew how to *use* the concepts to solve the problems they were given in class, but weren't sure *what* they represented. Consider the following excerpts from discussions about rejection regions.

> **Excerpt: Interview 172**
> M: So, I'm trying to make connections and some of the things you were talking about were these things called rejection regions. [I: Um hum] And…um…accepting and rejecting, and then *p*-values…so can you talk to me a little bit about what you mean by acceptance region, or rejection regions?
>
> I: Right…um…well…when you're doing a hypothesis test I…you would take the…um…x bar minus the…um…mu that you're given, or the mean that you're given in the null hypothesis and then…divide it by…ah…the standard deviation divided by the, ah, the number or the n which is the sample size. So…and…I mean, in this it's a large, ah…population so you would use…ah…*z*-values or…and…I…you would get a certain statistic but…when…because you could do a 0.05 ah…level or that's what the alpha equals then um…you have to look at the *z*-value of 0.05 and because in this case your alternative hypothesis is…that the mean is greater than 110 then your rejection region would be…um…the *z*-value greater than whatever…um…it is at 0.05. So, if…the test statistic you ended up coming up with…um…is actually greater than whatever *z* of 0.05 is then you reject the null hypothesis in favor of the alternative which…in this case is…that the mean IQ is greater than 110. Um…and then, the *p*-value from what I remember it to be was the…like, smallest number you could choose and then still…reject the null hypothesis….so you would take the probability of whatever your test statistic was in this case and it would also be *z* greater than or equal to that number…and then…to find it you would have to do 1 minus *z* less than or equal that number.

M: Ok (both laugh). I'm not sure I got all this. Ok, let's just back track [I: Ok] to the *z*-value. Um…so you did this kind of calculation, if I got it right [I: Um hum] x bar minus the null mu and then divided by the standard deviation divided by n [I: Yeah]. So…do you know why you do all of this?

I: Well…I don't know exactly why you put it in that form….but….I mean, that's the way that we had been taught for a while, or that type of form to…um…test for confidence intervals or test for…ah…like an something approximated to the normal distribution…so…I guess that that's what that's trying to do is to approximate this test to the normal distribution so you're able to test it cause you can actually find values then from the table…um.

Excerpt: Interview 192

M: …can you explain to me this rejection region and confidence interval is that you keep talking about?

I: (Laughs) Oh…um…I can't define them. I just know how to do 'em. (Both laugh) [M: Ok] Um…the confidence interval, when we do the calculation, if the number falls in the confidence inter, interval, it's more likely that…that it…it's accepted. I think we did that before we did the null hypothesis, or whatever. But, that (inaudible) and then the rejection we would do…we would use the alpha to come up with that…alpha, the degrees of freedom, and all that jazz. And, we would…um…if the number falls in the rejection region then the null hypothesis is rejected.

M: So what does that mean, the rejection region? How does that…?

I: They would…it's, um…it would be like…we need to do a calculation, come up with a number that…and the…and based on things that we learned in class like…ah…a whole bunch of things like with the *z* affective, I believe, he called it…do you know, of course you know (laughs). *Z*…um…with the…yeah when we come up with *z* and if the null hypothesis is that *z* alpha is…wish I had my notes (laughs) …it like, I think that (inaudible). *Z* alpha is…wait, wait, wait. If the null hypothesis is that mu is equal to…equal to….mu 2, I'm just going to use that number because I'm not sure what it, what it actually is…then, *z* alpha, the rejection region is that the absolute value of *z* is less than or equal to…a number (laughs) that we would ultimately find when we look it up in the chart and whatnot.

M: Ok.

I: So that would be the rejection region using the number. So, if it's greater than or less than or equal to, whatever (number). If it's…ah…falls within that, then you're supposed to reject it.

Excerpt: Interview 169

M: What is…what is this rejection region you keep talking about?

I: It's…ah…threshold for which your second hypothesis either crosses or doesn't [M: Ok] and when it crosses…and this actually for a while confused me because it seems a little counterintuitive to call it a rejection region. I was like, oh ok, well, if it gets above this number or if crosses that threshold then you reject the second one. But, it's actually

> not. When it crosses that threshold you accept your second hypothesis and then reject your initial hypothesis, your null hypothesis.
>
> M: If it crosses the threshold, you reject and if not…
>
> I: And, if not, you accept… you retain H0.
>
> M: So…if what crosses?
>
> I: Like a test statistic.
>
> M: And, what is a test statistic?
>
> I: It is your mathematical…it is your number that you derive from…ah…manipulating your sample mean and standard deviation as well as your purported…um…sample mean…and usually that's what you're testing, your sample mean. Um…you manipulate those three variables and then your sample size, which is usually *n*…um…and you add 'em up, subtract and do whatever you need to do...and, if the number crosses ah…the threshold whatever, of your rejection region, whatever that might be…if it's greater than some number or less than some number or the absolute value is greater than some number you…and then…and obviously you apply those three types of rejection regions in different, for different reasons and in different types of problems…ah, if it crossed that threshold then you make the determination.
>
> M: And, where does this threshold come from? [I: Um] What does that mean?
>
> I: Um…it's…it's based on…your…I believe the symbol is alpha…where you're given, it says test such and such at alph, alpha equals…we'll say 0.1 and then that means to me…divide that number, divide 0.1 by 2 and that gets your alpha divided by 2, alpha *z* divided by 2 number…for *z* alpha divided by 2, I'm getting my variables mixed up…and, ah, look it up in your *z*-chart if that's appropriate or your *t*-chart for smaller samples…and you get this number, whatever it is and…your rejection region…ah…you set it equal…depending on what you're second hypothesis is in which if the second hypothesis is asking for a greater than…greater than or equal to, a less than or equal to, or a does not equal…you change your rejection region's symbol based on that. And, your rejection region's simply…measuring one number versus another and seeing if it falls above that, below that…if it fits the criteria that you set out, that you set forth for it.

These examples are representative of the explanations that the interviewees gave for rejection regions. Again, we see that the interviewees describe a set of rules and/or openly express that they do not know exactly what the regions truly represent. There was no mention of the fact that these rejection regions are determined by referring to a sampling distribution of the sample statistic conditioned on the null. Nor did anyone mention that the rejection region provides a means of determining whether the sample collected in the study represents an unusual event conditioned on the null. Instead, the explanations offered by the students were very rule driven with a focus on procedures.

Additional attempts were made to probe student understanding of the logic and reasoning. As the interviewees mentioned *z*-scores and *z*- or *t*-tables in their explanations of rejection regions, they were asked to explain what these objects represented. These probing questions provided the interviewees with more opportunities to make the connection between the procedures and the underlying logic and reasoning. Unfortunately, this did not happen. Some examples of the interviewees' answers to questions about these concepts are included here.

Excerpt: Interview 81
M: Ok. So what do you mean by alpha level?

I: Um…this test…I don't really. It's the alpha. It just, like I, it just goes into the equation to test things. I really, I don't honestly know what it even means. I don't think I was ever told what it truly means. And, if I was I…I kind of just…I kind of just learn the equations (both laugh).

M: So can you explain at all, like, maybe not what it means but how it's used?

I: You find it on the chart like you…(writing) you have like alpha and then you have to find, like, 1 minus alpha and then alpha over two and then it goes and then you find that on the *t*-chart. And, that's a value you plug into the equation. What I…honestly just really not sure other than it's in the equation.

M: So you do alpha, one minus alpha, and you do alpha over 2?

I: Yeah….

M; And, you're putting it into a *t*-chart. What do you mean by a *t*-chart?

I: It's a *t*-table. It's some table we got that once you found that what alpha equals you looked it up and you found, I guess what *z* was…and, you put it into the equation you use to figure some of this stuff out.

M: So do you remember what that…*t*-table was telling you, or what those numbers all meant?

I: No…I don't…I…either I wasn't really told what they truly meant, or I didn't really note it down. Because, honestly I was just trying to remember how to do problems to get through the class…so…I don't really remember in all honesty. Sorry.

Excerpt: Interview 172
M: Ok. So then you look this up on, cause alpha's 0.05, then you said you look up the *z*-value of 0.05…and you're looking this up on tables [I: Right] I assume. Ok, so what do those tables…represent, like, what are the all these numbers that you see in the tables?

I: Well, they represent…um…well…the table…(draws) I guess, it, the all the table numbers are set up in this type of form…like the *z*-number less than or equal to some little *z*. And, the table gives you…um…across…ah, like a…0, 1, 2, 3,…and all the way up to about, like, 3.5 and then you can go across and find…ah…like 1 if this was like…1.1 you could find like 1.11 and if that's the little *z* that you're trying to look up then it gives you…um…the…ah…I think it, this is the probability of this? So this gives you the probability that *z* is less than 1.11 and the 1.11 refers to the place on the actual graph…the…bell curve (draws) where…um…so like this would be 1.11 on here…and…this number that you actually look up in the chart refers to…ah…the probability that it's less than that…so. And…I mean, before we did hypoth, hypothesis testing we were kind of finding…ah…*z*-values in terms of confidence intervals for…you could be sure…you know, 95% sure what the mean looking up, ah, where on the chart in here the probability is…if you were 95% sure you'd be finding 0.05 or as close to it as you could on the chart. And, then, seeing what little *z* that was.

Excerpt: Interview 169
M: (Writing)…of the rejection region. Ok, so what are…can you talk to me a little about what these *z*-scores are, this table that you're looking up? What is that telling you? What's the table show?

I: Um…it always, I far as I can say, it determine…it's ah…the distribution…it shows a distribution of…it reflects the distribution of your population. I was always really fuzzy about this, me and my friends always…me and my one friend from class were like…I really don't know, some of the time, what the differences are…but I know when to use which one. I can, you know, there are like small tells that you know to use a different table.

M: But you don't know what those numbers are representative of?

I: I'm thinking that it…I'm…I'm trying to say that it is…the probability that a number falls…ah, is less than on, on the *z*-chart itself, that a number is…a given number is less than…some other given number.

M: Do you know what those numbers are?

I: I don't think I do…I keep thinking…I keep just going it's either in the chart or you start out of the chart and work your way in.

Excerpt: Interview 122
I: So, like, I understand how they get the numbers, I just, um, will follow their concept. (both laugh) So I understand how they get everything. It just – I don't know how they come up with like, establish everything, but I understand by using the tables, like this SNV table, the chi-square table, like, I understand how they try to get these numbers.

M: Which also brings me to another question, then, these z-scores, what do you mean by a z-score?

I: Z-scores is, um, usually there's like an equation regarding to that, like, to determine, if I remember it, x minus…(writing)… er…*x* square over standard deviation, like if you give, um, if you're trying to find like a member of some sort, it's kind of like difficult to explain because I don't have my tables with me. This like – It's just telling you there's like a huge table. And then there's like this area of all these numbers. And that's your probability. The *z*-score, just is – just like a address kind of to place where your number is at.

M: Okay.

I: To see where your rejection area is at.

M: And then what do those probabilities for? They're probability of what?

I: Like, probability of some – something to occur, like try and do the hypothesis testing was a probability in 95% of this occurring. So we're trying to use that 95% con – change that to a decimal to 0.95 and then look at the – the SND table which is the standard normal, uh, distribution table to determine what is your z-s – well, your z-score to get this number.

M: Okay.

I: And like, by doing that, like you subtract the 1 minus 0.95 and you get a 0.05. And then if you find this number in the area, you – there's like – this is kind of difficult without the table, like you have these – you have numbers here [refers to her drawing of the table] and then you have numbers here and then you have the probabilities, the different number of probabilities you'd look was close enough to this, and then you had to find the other the go across to your, what is that, vertical to get that number And then you go to horizontal and get that number and add them together to get this.

Excerpt: Interview 192

M: Ok. So do you know what these *z* things come from?

I: The…(laughs)…honestly, no, because I've been trying to figure that out since we learned them…where…the…where we can find them in the chart. To a point where I just gave up (laugh) and used the numbers they gave us in…um…the *z* alpha numbers they, that they gave us in the book. So there's a little chart so you know if you have…when 1 minus alpha equals you know 99, 90% then ah…the *z* alpha over 2 is 1.945 or something.

M: And, was ever…were there any graphs associated with this at all?

I: Well with…with finding *z* itself, yeah there's graphs in the back of the book. But, the *z* alpha…the *z* alpha over 2, I really don't…know. It's like I go to class and he starts talking about it and they're like oh it's equal to this, it's equal to that and I'm like how do you find it? And, they're like look in the chart. And, I'm trying to look in the chart and I don't see it. So I just went to the book, opened the book to the chapter and

> there's a small little chart there. But then I would get questions about, you know…ah, percentiles that aren't in the book so in the book, like on the chart it goes from 80 to 99 and then I would get, you know what's the …80 to 99…but then it would go 80, 85, 90, 95, 97, 98, and 99 (laughs). Then, I would get, you know, what's the 96 confidence…get a…give me a confidence interval with a 96%.

Here, again, we see that the interviewees did not have a strong understanding of a concept important to the overall logic and reasoning of statistical hypothesis testing. The interviewees did not understand that the statistical tables give probabilities associated with the sample data under the assumed null condition. In some cases, the interviewees explicitly stated that they didn't know what the tables represented. In others, they offered very procedural explanations of the tables.

Another concept that came up in the interviews is that of the normal distribution. As evidenced by some of the excerpts presented above, some of the interviewees used graphs to explain their thinking about the item and/or to explain rejections and statistical tables. In the previous section (presentation of interviewee responses to the **Analysis of the Sample** and **Decision Rule** items) we saw that the interviewees do not understand what the graphs represent. Nevertheless, some of the interviewees referred to normal curves to explain their thinking for this item. These interviewees talked about the assumption of normality that is associated with statistical hypothesis testing. However, this discussion was often rule-bound as evidenced by the following examples.

> **Excerpt: Interview 192**
> M: So you don't know what this graph is a curve of…what that is?
>
> I: No. I just know it's a normal distribution. Right? But, that…cause that's what he tells me (laughs).
>
> M: So you don't…do you know what normal means?
>
> I: (Laughs)
>
> M: No.

I: (Laughs) I just do the graph. (Both laugh)

M: Ok…um…and do you…so then do you know why it's that shape and not some kind of other thing that would be…?

I: No

M: Like a parabola or a line or anything like that?

I: No. I just, they just tell me that normal distribution, shaped like a bow and I work with it.

Excerpt: Interview 172
M: Ok (both laugh). I'm not sure I got all this. Ok, let's just back track [I: Ok] to the *z*-value. Um…so you did this kind of calculation, if I got it right [I: Um hum] x bar minus the "null mu" and then divided by the standard deviation divided by n [I: Yeah]. So…do you know why you do all of this?

I: Well…I don't know exactly why you put it in that form….but….I mean, that's the way that we had been taught for a while, or that type of form to…um…test for confidence intervals or test for…ah…like an something approximated to the normal distribution…so…I guess that that's what that's trying to do is to approximate this test to the normal distribution so you're able to test it cause you can actually find values then from the table…um.

M: Ok, so…this gives you a *z*-score, right?

I: Right.

M: Ok. And, then, so you're saying that you think that this is to approximate it to the normal distribution.

I: Um hum.

M: And, why the normal distribution?

I: Because the sample in this case at least is…in the class we've been using that you approximate it to the normal distribution if your sample size is greater than 30. So, in this case, it's a sample of 500 students…um…so that's definitely greater than 30 so you would use the normal distribution as opposed to the *t*-distribution for samples that are smaller than 30.

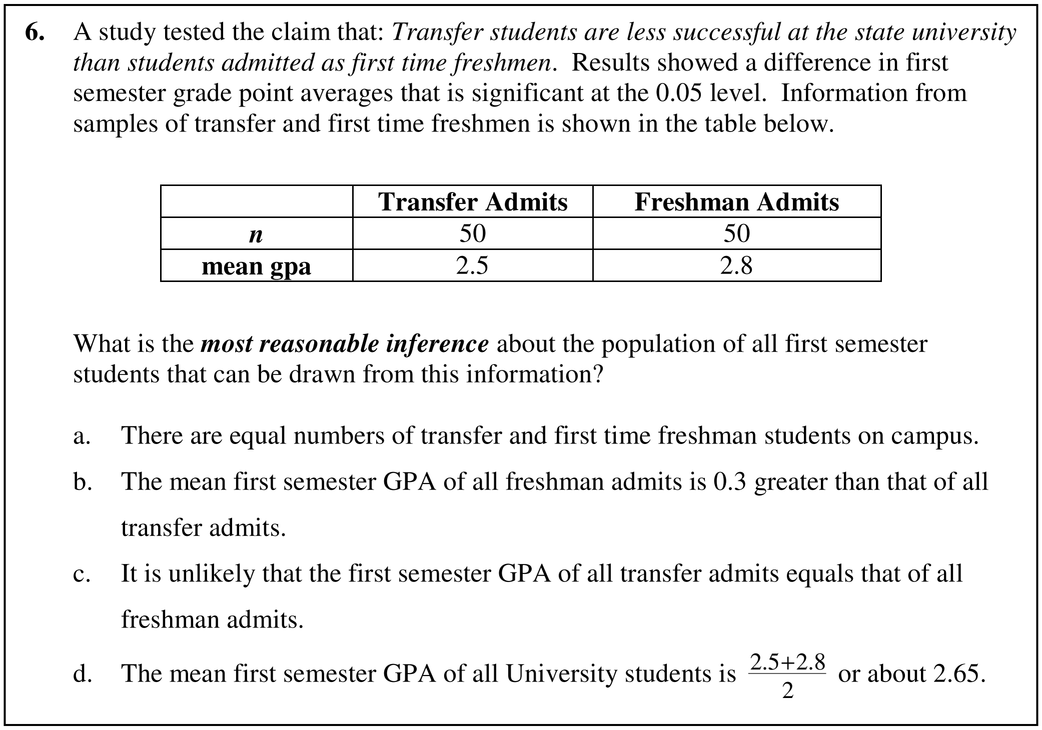
In addition to a "rule-bound" understanding of the connection of the normal distribution to the logic and reasoning of statistical hypothesis testing, the interviewees also did not have a strong understanding of *what* was assumed to be normally distributed. Many of the interviewees explained that the values in the sample were normally distributed, not the sample statistics for all samples of a given size.

In summary, a great deal of time was spent probing interviewees to not only explain their thinking about this item specifically, but also to explain their understanding of the overall logic of statistical hypothesis testing, in general. As a result, the discussion surrounding item number 5 was very useful in gaining insight into student understanding of the overall logic and reasoning associated with statistical hypothesis testing. Unfortunately, however, the discussion revealed that the interviewees did not have deep, connected understandings of the logic and reasoning associated with statistical hypothesis testing. They did not understand the test to be an analysis of the degree to which the sample collected is unusual conditioned on the null hypothesis. Hence, the interviewees did not interpret a statistically significant result correctly.

One final item from the **Recognizing Applicability** and **Conclusion** grouping was included in the follow-up interview and was used to further investigate the claims that introductory statistics students make when given a conclusion and the sample data. Classified in the **Conclusion** category, item number 6 assesses whether, given the results of a statistical hypothesis test, students understand what kinds of claims about that information are valid and which are not.

Figure 5.6

Item Number 6, Multiple-Choice Assessment



6. A study tested the claim that: *Transfer students are less successful at the state university than students admitted as first time freshmen*. Results showed a difference in first semester grade point averages that is significant at the 0.05 level. Information from samples of transfer and first time freshmen is shown in the table below.

	Transfer Admits	**Freshman Admits**
n	50	50
mean gpa	2.5	2.8

What is the ***most reasonable inference*** about the population of all first semester students that can be drawn from this information?

a. There are equal numbers of transfer and first time freshman students on campus.

b. The mean first semester GPA of all freshman admits is 0.3 greater than that of all transfer admits.

c. It is unlikely that the first semester GPA of all transfer admits equals that of all freshman admits.

d. The mean first semester GPA of all University students is $\frac{2.5+2.8}{2}$ or about 2.65.

This item is similar to that of number 5. However different information is given and the answer choices address different aspects of what can be inferred from a conclusion statement. The answer choices in this item address inference from a sample to the population. Answer choice *c* is correct. Answer choices *a*, *b*, and *d* are incorrect inferences because they are too conclusive. Each of *a*, *b*, and *d* make correct claims about the *sample* but these claims do not necessarily hold for the *population*. Even though the difference in GPAs was found to be significant at a significance level of 0.05, this does not mean that the sample statistics provide direct measures of the population parameters.

In the multiple-choice assessment, 7.7% of the participants chose answer choice *a*, 34.6% chose *b*, 36.5% chose *c*, and 21.2% chose *d*. Thus, performance was not very strong on this item. Based on the percentage of students who chose answer choices *b* and *d*, it was concluded that introductory statistics students often think that sample statistics provide direct measures of population parameters.

Of the eleven interviewees, 1 person chose *a*, 2 people chose *b*, 5 people chose *c*, and 3 people chose *d*. Over the course of the interviews, however, several people changed their answers. One person changed from *a* to *b*, 2 people from *d* to *c*, and 1 person from *c* to *d*. With these changes, no one had chosen *a*, 3 people chose *b*, 6 people chose *c*, and 2 people chose *d*.

Interviewee Explanations: A Summary

Over half of the interviewees chose answer choice *c* and used similar reasoning to justify their choice. They eliminated option *a* because they did not it was reasonable to assume that just because the sample size was 50, that meant that each of the populations also had 50 members. They eliminated answer choice *b* because it was too conclusive. The interviewees felt that it was not reasonable to assume that, just because the sample means showed a difference of 0.3 points, the population means would necessarily differ by 0.3. Many of them said the same for answer choice *d*, claiming it was too conclusive. Examples of this reasoning are presented in the excerpts presented below.

Excerpt: Interview 132
I: So…(reads out loud, somewhat inaudible)…hmm. Um…let's see, I don't…I don't know. None of them really seem that…ah…clear, now that I think about it. Um…I chose *c* but it…um…I guess I just looked at from like 2.5 and 2.8, they seem so similar that they're probably isn't a difference. But, that was just by like looking at numbers. *A*…um…I didn't consider correct because you're just taking a small sample of just 50 students and so that doesn't really even talk about how many students applied that were transfers and how many students who applied that were freshmen…so. And…and I

guess *b* and *c* are both…they both seem unlikely to happen because this is just you're taking 50 students who are transfers and freshmen, and so…yeah, the freshmen do have a 0.3 higher GPA but ah…it says that all freshmen have a higher GPA than all transfer students…and, so. That didn't really seem very likely.

M: Ok. And what was *d*?

I: *D*…ah…for the same reasons. They're talking about all university students…ah…so that means seniors, juniors, sophomores, and freshmen. And, none of that information is given, and so. Yeah. Oh yeah I said all students and so, but here it's just freshmen and transfer students and so that prob…that doesn't represent all the students in the college.

Excerpt: Interview 15

I: Well, for like *a* in 6…like you didn't…um, you didn't test all of them so you can't say that um, the whole populations are equal. You just had equal sample numbers.

M: Um hum.

I: And like again you can't make a concrete conclusion about all of them, because you didn't test all of them, you only tested a sample.

M: Is that for *b*? (inaudible)

I: Yeah, so *b* would be true for your sample but you can't like really say that it's for all of them. Cause, that may not true…like, if you go back and test all of them, it might be different. And then…like for *c*, that's your sample found that…like they're not equal. So you can probably say like based on your sample that it's unlikely for all of them. Because you're not like going out there and making like a concrete statement about all of them, because you can't possibly know that because you didn't test all of them. So saying it's unlikely means that like you're pretty sure but, you know, there's a possibility that you're wrong. And, like, *d* again is like making a statement about all of them.

M: Does the fact that they did this test up in the beginning, the stem…and they found it to be significant at the 0.05 level, does that enter into anything that you're answering there? Your thinking about these?

I: Um…I guess because it's significant like you can …um…say that like since the population like this…that you can…I guess you can say that freshmen GPA's greater than the transfer for your sample. And, so, you can say like they're unequal and then so you can project some of that conclusion to the population.

Excerpt: Interview 169

I: I chose *d* but that's not…that's not what I would choose now. I must have just completely read *c* wrong, because I would now choose *c*.

M: Ok. And, why?

I: Um…because…(inaudible)…well it's conjecturing that the study…the study's claiming that transfer students are less successful than first time freshmen and so…that means to me, that tells me, that helps me…ah…made…you know, make obvious that transfers are a small group out of a large group…that transfers versus all freshmen. And, the assumption being that it's, there's only two groups in the large group of freshmen students. that transfers are one and…ah…everyone else is another. And, so…it's claiming that…my first impressions of the data…lead me, lead me to believe that that's right. And, they're and *c* fits that…*c* fits that criteria…or *c* wait, no, *c* helps me put that, verbalize that. *D* is not right (laughs). I don't know why I chose it.

M: Why isn't *d* right?

I: Um…it's asking for…it's trying to make a claim for all University students…when it's not…it doesn't take into account…when it says all university students that means…freshmen through graduate students, freshman through seniors. It's not the right…and not only that, you can't know the mean based on a small sample, like that. Or, I mean, a small sample consider, you know, under that assumption that a University has a lot of…people at it.

M: So, if it would have said that the mean of, mean first semester GPA of fresh, freshman and transfer students is that…would that be correct? Or still not correct?

I: It's making its way toward correct but it's still not. Um…

discussion continues

I: *A* is not right…um, it just, the sample just happened to be the same size. And, again, *b* makes a very general claim. I mean the whole point of hypothesis testing is that you can't test an entire population.

Excerpt: Interview 112

I: (reads aloud, somewhat inaudible) Yeah, you can't assume there's equal number of freshmen and transfer students on campus just because the sample size is the same. So, that's why *a* can be thrown out. (reads *b* out loud, somewhat inaudible) You can't assume… that, that for all either. Cause that hasn't been proven beyond the 0.05 significance level. (reads out loud, somewhat inaudible)

M: Oh, wait. It did say that it was.

I: Well…well, yeah, it is significant at the 0.05 level, but only there. Yeah, it doesn't go beyond that…necessarily. Like it might not be significant at the 0.01 level…or like the 0.001 level.

M: Oh, ok.

I: So you can't assume that it necessarily has to be.

M: Ok, so it may not be at a smaller level

I: Yeah

M: (writes)… be significant at a smaller level, so it can't assume it's true for all. Ok.

I: (reads *c* out loud) Yeah, I figure that…I figured that one was the best because they're…cause at the significance level, they're not equal. And, it…like…I don't think it's like a good enough significance level like you can…can assume that that is the…like difference there but at least you can know that there…there is like a… it…it is higher for the freshman admits. It's fairly significantly known that. And, then *d* the mean first semester GPA of all university students is…yeah, I thought… that just didn't seem right to me adding them together and getting an average using that.

M: Why not?

I: Hmm…I thought that…hmm…I don't know (laughs). I thought they should probably have to like add out all of them from the individuals and then divide it. I could be wrong, though.

M: All the individuals….?

I: And anyway it's only…anyway it's only at the 0.05 confidence level too.

Others eliminated *d* because they were not sure whether the calculation of the mean was correct. Some excerpts that illustrate this reasoning are included below.

Excerpt: Interview 77
I: Ok. Ok, well *a* is kind of obvious that there's equal numbers of transfer….no…wait. Well *a*'s not true because these numbers were from the sample. It's not saying that there's 50 transfer admits and 50 freshmen admits. And, so *a* would not be true cause it's saying that there are equal number of transfer and first time freshmen students on campus, since these numbers are just taken from…these are sample from a larger population.

M: So you can't just [I: Yeah] believe that that's going to be the case?

I: Yeah. Um…I …I mean, I would say, *c* because it's showing that the mean GPA between the transfer and the freshmen students, um, is different…between the samples that were taken. And, so you could say that it's unlikely that the GPA in the two…two categories of students are equal. Um…*b*…I …didn't think would be the most reasonable inference because this is just a sample that was taken so, the number where it's saying is 0.3 greater…um…that could be much higher or lower within the whole population. I'm not really sure about *d*, um…I don't think it's right…or I don't think that you could just say that the mean first semester GPA of all university students is just the average of the two…um. I mean it could be…but…

Excerpt: Interview 29
I: Maybe I just didn't read. (Both laugh) Um…I guess *c* could be true (laughs). Now, I guess I change my answer to *c* (laughs).

M: Ok. So now why don't you like *d*…or *a* or *b*?

I: Um…(reading *a* out loud) there are equal numbers of transfer and first time freshman students…like, I don't know. It seems…kind of…I mean that's just like the size of the sample, so…I mean, you don't know how many like transfers or freshmen are actually on campus. [M: Ok] And, um…and like because of the small sample you don't know…I don't know…like…I guess because like I thought the sample was like…it was only like 50 people, or 50 of each…so, I guess to say…that the mean is 0.3 greater than that ….it's just….it is…too conclusive (laughs)? Cause, I mean, you're actually sure for certain so, I don't know, yeah (laughs).

discussion continues

M: Ok. And, why not *d* now? You switched to *c* now.

I: Um…I guess *c* just seems like a better answer now (laughs). I'm not really sure about *d*. Like I don't know whether like taking the average of two means, like, if that would…um…you get a better average, I guess, of….the two groups.

These excerpts illustrate that, on the whole, the reasoning used by the interviewees who chose the correct answer was similar. It is interesting to note, however, that in their reasoning, only a few of these interviewees referenced the fact that the results were found to be statistically significant. Answer choice *c* is legitimate, especially since we are told that the results were found to be statistically significant at a level of $\alpha = 0.05$. With the exception of Interviewee 112, none of the interviewees mentioned this fact. Based on their responses, it is not clear whether they chose *c* merely because the sample data supported the notion that the two GPA's are not equal (as did Interviewee 132) or whether they did so because the results were significant and they simply neglected to mention it in their reasoning. Nevertheless, they all understood that *a*, *b*, and *d* were not appropriate inferences.

Two interviewees chose answer choice *d*. Interviewee number 172 thought both *c* and *d* were reasonable inferences but she was more comfortable with answer choice *d* because she remembered doing the calculation suggested in option *d*. Her reasoning is illustrated in the following excerpt:

Excerpt: Interview 172

I: Ok. (Laughs) Um…ok, well I chose *d* in this case at least I think just because…um…I knew that…ah…this is how you could find the, ah, the mean from

> examples that we've done in class where they just give you information like this and then it asks you for the mean and standard deviation. Or, the, not the standard deviation, the standard error. Um…so I knew that that was how you could find the mean of all the…ah…students and, um….
>
> *discussion continues*
>
> I: …And, then, also because it says there are equal numbers. I mean, in this case, this is just a sample of both of them, so…and you would want your samples to be equal to be able to…um…perform hypothesis tests. And…the same thing with *b* that it's just like because these are both just samples…um…you don't know if like….in general you can make the assumption that all freshman admits have the 0.3 greater GPA than transfers. And, um…I mean, I guess….*c* is a…reasonable…inference but it's just…um…I don't know I was more sure about *d* because I knew that that was a way that you could calculate the mean from the things that were given.
>
> M: Ok. So again, this is one of those what's the most. So it sounds like to me that you knocked out *a* and *b* and said they're not [I: Um hum] even possible. But then *c* and *d* are possible but *d* you felt more comfortable with because you've done those kind of calculations before [I: Right]. Ok.

Interviewee 172 was used to performing the calculation described in answer choice *d*. Therefore, she chose that option. She did, however, recognize that answer choice *c* was also a valid option.

Interviewee 192 also chose answer choice *d*, but not until she had struggled to decide whether answer choice *c* was correct. She had initially picked option *c* but then, over the course of the discussion about item number 6, changed her mind. Her reasoning is illustrated in the following excerpt:

> **Excerpt: Interview 192**
> I: (Reads number 6 out loud) So…mean GPA 2.5, freshmen admits 2.8…it is unlikely that the first semester GPA of all transfer admits equals that of all freshman admits…hmm…why did I chose that? Um…not looking…intuitively, I guess…because transfer…for the transfer admits…hmmm… (Reading *a*) There are equal numbers of transfer and first time freshman students on campus…I mean, *a* is… I think that I did it by elimination. I mean, *a* is not a…something that didn't make sense to me. There's no way there would be equal numbers of both.
>
> M: Even those these numbers (pointing to the 50 in both columns) are the same?
>
> I: I mean, it…I got…I mean…that I guess that's true, but…(reads *b* out loud, somewhat inaudibly). Hmm…why'd I pick *c*? I don't know why I picked *c*. Because, looking at *a*, *b*, and *c* they're all true based on the chart. And, then, *d*…would it be *d*?

> Because the *d* is something you logic, logically infer based on the chart because…that the mean GPA of all first semester…students…
>
> *discussion continues*
>
> I: I wanna choose *d*.
>
> M: *D*. Why?
>
> I: Because *a*, *b*, and *c*…you can get from the chart and *d* is something that you have to actually…ah…put together from the information in the chart.

Here we see that Interviewee 192 changed her answer to *d* because, for her, it was the only option that constituted an inference. She reasoned that the claims made in answer choices *a*, *b*, and *c* could all be justified directly from the table. However the claim made in answer choice *d* could not. A calculation was required to make that claim. Therefore, it is the only statement that was *inferred* from the table. Unfortunately a misunderstanding of the meaning of the word "inference" caused Interviewee 192 to change her answer from one that was correct, to one that was incorrect.

A misunderstanding of the word "inference" was also an issue for two other interviewees. Interviewees 81 chose answer choice *b* because he thought it was the strongest inference that could be made from the data. His reasoning is illustrated below.

> **Excerpt: Interview 81**
>
> I: Ok. Um… I said *b*, the mean first semester GPA of all freshmen admits is 0.3 greater than that of all transfer admits. And, the reason I said that is because it, it is…the mean GPA for freshmen admits is, ah, 2.8 and transfer is 2.5, which is .3 higher. Which is why I chose that cause it was basically directly from the data. I felt that was the best answer. Like, there, there were an equal number of transfer and first time freshmen students on campus didn't really seem to have much to do with it. Um…it's unlikely that the…well that was *a*. Ah, *b* was what I chose. *C* says it is unlikely that the first semester GPA of all transfer admits equals that of all freshman admits, which was true but I didn't really believe that the, ah, study was really about that. And, *d*, the mean first semester GPA of all University admits is 2.5 plus 2.8 over 2, or about 2.65, I mean, again that's true. Um…but I still felt *b* was the best answer because it was taken directly from the data like you could…obviously *c* there was a 0.3 difference in the, ah, GPAs.
>
> M: Ok. So, you said you didn't pick *c* because you didn't think that was what the study was about. So, can you say more …what you mean about that.

I: Um…it is unlikely that first…(reads out loud, somewhat inaudibly). I mean, it says that it's unlikely but I mean it's not really proving that because it wasn't asking you if it was likely they were the same. I mean 0.3 off is, isn't that much. So, I, that's technically true. Um…but I, I felt *b* was a more obvious and better answer because it, it was taken directly from the data. I didn't really think *c* had much to do with it, due to the fact that…um…the question was talking about inference. Even though it's true I just didn't think…I didn't think it was a strong of an inference as *b*.

Interviewee 81 chose answer choice *b* because it was information that "could be directly taken from the data". For him, that was what constituted an "inference".

Interviewee 122 also struggled with the meaning of the word inference. Consider the following excerpt from her interview.

Excerpt: Interview 122
I: Six? Okay. A study tested in the claim that transfer students are less successful at the state university than students admitted as first time freshman. The resu – results show the difference in first semester grade point average that is significant at the 0.05 level. Information from samples of transfer and first time freshman is shown in the table below. What is the most reasonable inference about the population of all first semester students that can be drawn from this information? Um…Okay, what I… what I got the answer to, um, a. There are equal numbers of transfer first time freshman students on campus. Um, this usually is, like in my opinion, or most of the time, like they have the same sample size so you could tell, um, that there is a great change or not, but then…but like, um, it kind of makes sense, because, um, I don't know how to explain it. Um, because it's easier to work with a sample size that have, if you have, ch… look at the transfer students that are 50 and the freshman 50, you could do a better comparison, because you have a, um, same amount of, um, like GPA, say like, uh, a odd number, like one is 43 and the other one is like 47.

M: Okay.

I: That's my reasoning for that one. Um, b, the mean for semester GPA of all freshmen admits is 0.3 greater than that of all transfer admits. Well, usually, um, regarding to make an inference, they don't really look at, because you're trying to find which actually was your samp, your sample size and your mean. Um, this will probably come later on regarding to trying to determine the, a reference. It's like the, the difference part, it'll probably be later on. Um, c, it is unlikely that the first semester GPA of all transfer students equal that of all freshman admits…well, it kind of, that kind of like, kind of seemed, kind of a bit biased in regarding to transfer students and freshman like ad, admits, because sometimes you never know that probably transfer students can do this as well as freshman students also. They just – this sample size, it just 50, so you can't really get a broad aspect regarding to that one. Um, d, the mean first semester GPA of all university students is 2.65. Hum, well I could use actually a and d. Okay. I think like I'd be able to, I don't know why I crossed it out, but I think I'd be able to use d because regarding to trying to find a, um, a statistical inference, you would need to, like, try to find a way to use the mean, median, mode and the mean is like the best way to determine a difference between the two population or see an average between the two.

> M: Okay. So…
>
> I: Because like mostly to me, like in my opinion, like the sample size and the mean…like, the GPAs are more important than like finding like subtracting those two because I can't really…I mean as much.

Interviewee 122 chose option a because, for her, the most reasonable inference was the part of the test that was the most important. She thought it was most important to have equal sample sizes, so she chose answer choice a. Hence, the discussion with Interviewee 122 is another example for which an interviewee did not understand the meaning of the word "inference" and chose the incorrect option. By the end of the discussion, however, Interviewee 122 changed her answer to choice b. But, it wasn't clear why she did so and, due to time constraints, there was not time to probe more into her thinking.

As was the case (with the exception of Interviewee 112) for the interviewees who chose the correct answer, we do not see evidence that the interviewees who chose the incorrect answers considered the fact that the difference was found to be statistically significant at the $\alpha = 0.05$ level in their thinking about this item. Thus, we are not sure whether the interviewees used that piece of information in their reasoning or if they simply ignored it. If the interviewees ignored that piece of information, the question remains as to how their reasoning may or may not have changed if they had paid attention to it. Fortunately, there was an opportunity to raise this issue with one person: Interviewee 191.

Interviewee 191 initially chose answer choice b but began to question his choice. In the interview, he was asked how the fact that the difference was found to be statistically significant at the α = 0.05 level of significance impacted his reasoning. The following excerpt illustrates his thinking with that additional piece of information.

Excerpt: Interview 191

I: Ok, so looking, just looking at the information and not the answers yet, um, obvious inference would be that this, this hypothesis or this claim, transfer students are less successful at the state university than students admitted as first time freshmen, um, the table shows that to be true.

M: Ok.

discussion continues

I: No…for…no, um, sorry, I'm getting confused with the population and the sample. All right, well you can't infer that based on – because these are only samples, so you can't infer that

M: Ok. Ok.

I: Um …the mean, or *c*, it is unlikely that the first semester GPA of all transfer admits equals that of all freshmen admits. It's unlikely that…I would say *c* isn't the most reasonable because, ah…well, I didn't…grades vary throughout everybody. Um, it's unlikely…so I don't think this conveys…that's asking about individual grades of students. Um, you can't really infer anything about the individual grades, I don't think. Wait a minute. It is unlikely that the first semester GPA of all transfer admits equals that of all freshman admits. Ok, that, that is likely, but I don't think you use the table to find that out. Um, the…the mean first semester GPA of all university students is…um, um, *d* I wouldn't use because well, you don't know the size of the university, so you don't know how valid these samples are. It could be a very small percentage or it could be a very large percentage. You don't know that for sure.

M: Ok.

I: So that's why, um, I wouldn't choose *d*, ah, and so yeah, I chose *b*. Ah, I chose *b*, the mean first semester GPA of all freshman admits is 0.3, greater than that of all transfer admits. Well…well, actually that's kind of, looking back at *b*, it's asking, ah, you to assume something about every student's individual grade…

M: Ok, the question asks for all – most reasonable inference for all – with all the information, so not just the table, but also the fact that they did the test and it was – the differences were found to be significant at the .05 level.

I: Mm hmm.

M: So they already did everything, and this was just their sample information, so taking that altogether, I'm hearing you say that *a* and *d* are not inferences that could be made. Am I correct?

I: Um, yeah.

> M: You shouldn't make those inferences. *B* you chose, but now you're thinking that you shouldn't make the inference either.
>
> I: Well, that – when I was – when I was looking at that though, again, you mentioned that, um, it's proven significant at the .05 level. Um, so I think considering that, that means that this information you can be pretty confident in, so I think they're saying with that, you can assume that, ah, the average – if you take a student, um, a freshman student, you can assume that his – his or her grade is somewhere around a 2.8, um, and you can assume with transfer students that their grade is going to be about a 2.5, and you – with – because it's significant at the .05 level, you could be confident in those assumptions, I think.

Immediately after Interviewee 191 reread the item, he made a statement that sounded very much like the claim made in answer choice *c*. However, Interviewee 191 noticed that he chose answer choice *b*. As he talked through the options, he began to question his choice. He was concerned that the claim in answer choice *b* was too conclusive. After it was pointed out that the difference was found to be significant at the 0.05 level of significance, however, he convinced himself that *b* was, indeed, the best option. Because 0.05 is low, one can assume that the sample statistics provide (fairly) direct measures of the population parameters. He reasoned that this direct translation of measures was reasonable because the results were found to be significant at a low level of significance.

Based on these results, it is reasonable to ask whether the same reasoning would have been employed by the other interviewees had their attention been drawn to the fact that the difference was statistically significant. It may be, however, that the interviewees did consider this piece of information in their reasoning. Nevertheless, there seems to be a general consensus that the sample data is just that. It is data from a sample. The population may or may have the exact same parameters as the sample statistics.

Analysis

Overall, the explanations provided by the interviewees for the items from the **Recognizing Applicability** and **Conclusion** categories indicated that they did not have strong understandings of the role of inference in statistical hypothesis testing. They did not appreciate the value of the method as an inferential tool, nor did they understand the nature of the inference drawn by statistical hypothesis tests. They were uncomfortable with the words “uncertainty” and “probability” and were quick to make strong claims about populations based on information from a sample and/or based on the results of a statistical hypothesis test. While the interviewees did not believe statistical hypothesis tests to be methods of proof, they did believe that statistical hypothesis tests provide information on the degree to which a given hypothesis is true. In addition, given sample information and a statistically significant result, the interviewees believed that there is a strong chance that the sample statistics provide direct measures of the population.

The follow-up interviews provided additional insight into student thinking about the overall logic and reasoning of statistical hypothesis testing. Based on their explanations for their answer choices, it was clear that the interviewees did not have connected, complete understandings of the logic and reasoning of statistical hypothesis testing and that this lack of understanding impacted the way in which they drew inferences about sample information and the conclusions offered by statistical hypothesis tests. The interviewees did not understand the role of indirect reasoning in the method and, therefore, did not have a strong understanding of what it means to say that a result is statistically significant. In addition, they did not understand the role of probability and sampling distributions in determining whether a result is statistically significant. When asked to explain these concepts, the interviewees recited rules and procedures. Such responses indicated that the interviewees' understanding was not well developed. Analyses of student responses to further questioning indicated that the interviewees did not understand the reasoning and logic that supports the procedures they were describing. They did not understand how those procedures are used to determine if a sample is unusual conditioned on the null hypothesis. Unfortunately, this gap in understanding impacted their overall reasoning about the results of a statistical hypothesis test and how those results are valuable.

Summary

The data and analysis in this qualitative phase of the study confirms what was found in quantitative phase: introductory statistics students do not have strong understandings of the logic and reasoning that support statistical hypothesis testing. The data collected in this phase indicates though introductory statistics students can perform the operations involved in statistical hypothesis testing, they do not have a deep, connected understanding of the logic, reasoning, and concepts that support those procedures. Analysis of interviewees' explanations for choosing the answers they chose on the multiple-choice assessment indicated that students have disconnected, incomplete understandings of the logic and reasoning of statistical hypothesis testing. In addition, the data indicates that introductory statistics students struggle to make those connections and, thus, are not able to articulate their thinking very well. These results extend conclusions from the quantitative phase as they pinpoint more specific components of statistical hypothesis testing that introductory statistics students do and do not understand.

The results of the qualitative phase indicate that introductory statistics students do not understand that statistical hypothesis tests employ indirect logic in testing to see if the sample presents an unusual case, conditioned on the null hypothesis. They do not have a conceptual understanding of sampling distributions. They do not understand what the graphs of sampling distributions represent, nor do they understand the reasoning that supports their use in statistical hypothesis testing. In particular, introductory statistics students do not have a strong understanding of the role of probability in the logic and reasoning. In fact, many of them do not associate the words “probability” and “uncertainty” with statistical hypothesis testing. They do not understand that a result is statistically significant if the probability of obtaining the collected sample, conditioned on the null, is low. They are able to use statistical tables and the level of significance to determine whether or not to reject the null hypothesis, but they do not know that those tables give probabilities associated with sampling distributions for a given sample size conditioned on the null hypothesis. They do not understand that a rejection of the null hypothesis is an indication that the sample is unusual.

Furthermore, the results indicate that because introductory statistics students do not understand the overall logic and reasoning associated with statistical hypothesis testing, they do not have strong understandings of the role of inference. They often make inferences about the population directly from the sample data, rather than noting whether or not the result was statistically significant to begin with. Hence, introductory statistics students do not understand the value of statistical hypothesis tests in providing a means of taking the variability that exists in sample data into consideration when drawing a conclusion about the population. In addition, many of the inferences that students make are invalid. They believe that statistical hypothesis tests, through the level of significance, provide a measure of the degree to which a given hypothesis is true or false. However, they don't know why they can or cannot make that claim. They do not have a strong understanding of statistical significance, how it is attained or of what inferences may be made based on it.

Though they do not understand the overall logic and reasoning of statistical hypothesis testing, the results of the qualitative phase indicate that introductory statistics students do understand that sample data is being used to determine whether or not one should "have faith" in a null or alternative hypothesis. They understand that the null and alternative hypotheses must be contradictory and that the alternative hypothesis represents that which the researcher would like to "prove". Introductory statistics students, however, do not have an understanding of *why* the alternative hypothesis represents that which the researcher would like to prove. When asked about this idea, they tend to cite rules, rather than connect it to the overall logic.

Finally, the results of the qualitative phase indicate that introductory statistics students understand that statistical hypothesis testing does not necessarily "prove" a hypothesis. This result is interesting in that it wasn't clear this was the case based on the results of the quantitative phase. Often, students chose answer choices that contained the word "prove". As a result of the follow-up interviews, however, we see that students often do not associate this word with its formal, mathematical definition. Based on the responses to the questions offered by the interviewees in the follow-up interview, it seems that introductory statistics students consider statistical hypothesis testing to be a "proof" in the sense that it provides an accepted way to justify a conclusion to a particular community. This additional information about student thinking helps to clarify some of the results obtained in the quantitative phase.

In summary, the results of the qualitative phase indicate that introductory statistics students do have a general idea about overall goal of statistical hypothesis testing. These students understand that the test is used to determine the validity of a given hypothesis. They also understand that the test is not a proof. However, introductory statistics students do not fully understand the logic and reasoning that supports the procedures that move a researcher from sample to conclusion. They do not understand the roles of inference, probability, and indirect reasoning in statistical hypothesis testing. Unfortunately, these ideas are the very essence of statistical hypothesis testing. Therefore, more must be done to help introductory statistics students develop stronger understandings of the logic and reasoning of statistical hypothesis testing.

Research Sub-Question Number 3

Results from the quantitative phase indicated that, overall, introductory statistics students have stronger understandings of the relationship between the context and the method than they do of the logic and reasoning associated with statistical hypothesis testing. However, these understandings are still fairly weak.

Analysis of the results of the ***method and context*** items multiple-choice assessment indicated that introductory statistics students know when statistical hypothesis testing can be applied to answer a question of interest. However, they do not have a strong sense of statistical hypothesis testing as an *inferential* method that relies on probabilistic concepts to deal with the variability associated with samples. Therefore, introductory statistics students do not have a strong understanding of the factors that should be taken into consideration when using and interpreting the results of a statistical hypothesis test within “real world” contexts.

As was the case for research sub-question 2, the results of the multiple-choice assessment only tell part of the story. It was, therefore, important to explore the issues raised in the quantitative phase in the follow-up interviews. By asking the interviewees about their reasoning for the items, it was possible to further understanding how students thought about these concepts and ideas.

In order to address the third research sub-question, three multiple-choice items were included in the interview (see Table 5.1). These items represent the **Recognizing Applicability**, **Collect a Sample**, and **Implication for Practice** categories from the Framework. For each category, a summary of student explanation will be presented and followed by an analysis of those responses. This analysis will be followed by a final, overall summary of the data and conclusions associated with research sub-question number 3.

Recognizing Applicability

From a ***method and context*** perspective, items in the **Recognizing Applicability** category were written to assess whether students understood the value of statistical hypothesis testing for answering questions about populations for which it is impossible or impractical to collect information from every member of those populations. These items assess the degree to which introductory statistics students recognize the conditions under which statistical hypothesis testing would be an appropriate, valuable method of investigation. Introductory statistics students should understand that in order for statistical hypothesis testing to be a useful method of investigation (1) the research question must address a well-defined population, (2) it must be possible to answer the research question using a measure of that population, and (3) it must be possible to construct two mutually exclusive, contradictory hypotheses to answer that research question. That is, introductory statistics students should understand that statistical hypothesis testing is a powerful *inferential* tool used to answer "measureable" research questions about large population and items classified in the **Recognizing Applicability** and ***method and context*** categories assess whether introductory statistics students have these understandings.

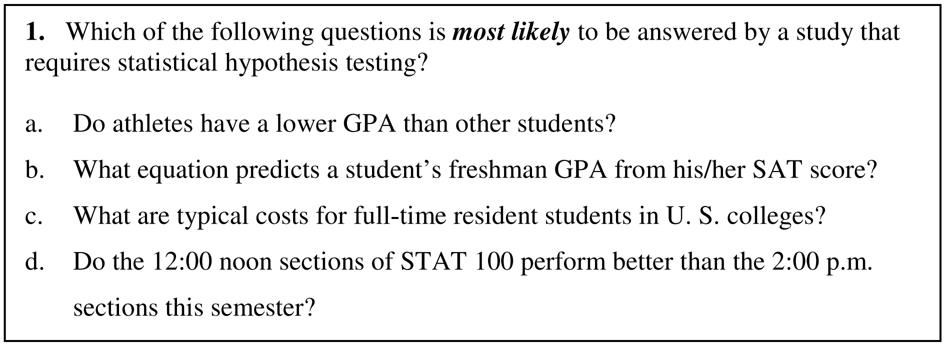
The results of the quantitative phase indicated that introductory statistics students understand that statistical hypothesis testing is useful for answering research questions in which two populations are compared. Furthermore, the results indicated that introductory statistics students understand that in order to compare populations using statistical hypothesis tests, it must be possible to define a measure of the populations that will be useful to answer the research question. They do not however, seem to understand that statistical hypothesis tests are valuable *inferential* methods and are useful when data on the entire population is impossible to collect. They do not understand that statistical hypothesis testing is not useful if information on the entire population can be collected.

These results give some indication of how students think about the value and uses of statistical hypothesis testing. However, data collected in the follow-up interview provides more insight into their thinking. Therefore, one item from the **Recognizing Applicability** category was included in the follow-up interview: Item number 1.

Item number 1 assesses whether, given a set of research questions, introductory statistics students are able to choose which question is appropriately addressed by statistical hypothesis testing.

Figure 5.7

Item Number 1, Multiple-Choice Assessment



1. Which of the following questions is ***most likely*** to be answered by a study that requires statistical hypothesis testing?

a. Do athletes have a lower GPA than other students?

b. What equation predicts a student's freshman GPA from his/her SAT score?

c. What are typical costs for full-time resident students in U. S. colleges?

d. Do the 12:00 noon sections of STAT 100 perform better than the 2:00 p.m. sections this semester?

Answer choice *a* is correct. The measure and populations are clearly defined and contradictory hypotheses exist that could answer the research question. The research questions offered in answer choices *b* and *c* do not meet the criteria for use in statistical hypothesis testing. Answer choice *d* is tempting distractor. The research question meets the criteria. However, the question can be answered without statistical hypothesis testing. Because the populations are small, there is no need to collect a sample. The question can be answered through direct comparison of the means for the entire populations described. There is no need for inference.

In the multiple-choice assessment, 43.3% of the participants chose *a*, 1.9% chose *b*, 2.9% chose *c*, and 51.9% chose *d*. These results indicate that students understand statistical hypothesis testing to be a means of comparing two groups but that they don't understand the role of inference in that test. This issue was explored further in the follow-up interviews.

Of the eleven interviewees, 4 chose *a* and 7 chose *d*. No one chose options *b* or *c*. Over the course of the interview, 2 people changed their answer from *d* to *a* so that, ultimately, 6 interviewees had chosen *a* and 5 had chosen *d*.

Interviewee Explanations: A Summary

As was indicated by the results reported above, none of the interviewees chose answer choices *b* or *c*. The explanations offered by the interviewees for eliminating these options were very similar. Many of the interviewees eliminated *b* and *c* because the research questions proposed in those answer choices were not questions that compared two groups. Some examples of such explanations are presented below:

Excerpt: Interview 192
I: Um…because….*b* doesn't really, I can't see a hypoth… I can't get a hypothesis, or a null hypothesis from *b*. And *c*, (reads *c* out loud, inaudibly). *C* asks a direct question. I'm not really comparing two hypotheses.

Excerpt: Interview 172
I: Um…well, I guess it's just because the way it was presented it seemed…most like the, ah, test that we had been doing in class but…um…I think I started looking at, ah…I knew, I didn't think it would be *b*…because…I don't know, I just…I couldn't think of a way to put it into that…um…type of formula, I guess. And…same thing with *c* because you weren't testing…ah…something against another thing. And…I also, I was kind of just, I came down to between *a* and *d*…but…I don't…I just thought that *d* fit it more than *a*, I guess.

Excerpt: Interview 77
M: Ok, why not *b* or *c*?

I: Um…because *b* is asking for just an equation to predict a…ah…that population GPA but that's not really testing whether it's lower or higher against another population. And, *c*, um…that also was just asking for like a certain type of data. It's not like comparing. It's not saying, you know, are the typical costs higher for full-time residents or like lower for…part-time…that kind of stuff. It's just asking for like a certain…certain quantities.

While the majority of interviewees cited a lack of comparison as a reason for eliminating answer choices *b* and *c*, there were a few interviewees who offered other justifications. Consider the following explanation offered by Interviewee 132.

> **Excerpt: Interview 132**
> I: Um…the other two ones, *b* and *c*, like you're talking about multiple equations so it couldn't be just one or…one equation or another. And, same with typical costs for full-time students in college. There are many different costs and so it would be hard to just have two hypotheses. You'd want multiple, I guess.

Here, Interviewee 132 did not focus on the lack of comparison in answer choices *b* and *c*. Instead, she was concerned that more than two hypotheses could be generated to answer the research question. Thus, she did not think these options were appropriate.

Other explanations were offered by Interviewees 112 and 122 who knew that statistical hypothesis tests did not find equations or average costs.

> **Excerpt: Interview 112**
> I: Well, *b* - what equation predicts a student's freshman GPA from his or her SAT score – I pretty much like…hypothesis testing doesn't find equations. So I ruled that one out immediately. And, then *c* was for the typical costs and I thought that was more of an average type thing. So I crossed that one out. And, then, I was kind of divided on *a* or *d*, because it was both two different groups. One was athletes versus the other students and the other one was like…sect…noon section versus 2 pm section. I mainly just chose *a* because it seemed like an example from the textbook.

> **Excerpt: Interview 122**
> I: Um, well, *b*, what equation predicts a student's freshman GPA from his or – SAT score? Um, well, it's kind of difficult to find an actual equation to be able to determine how a GPA ref – um, correlates with the SAT scores. And with C, what are typical costs for full-time residence students in U.S. colleges? Well, it depends, like, you can't really find, like, a statistical hypothesis because it varies from each person because there is financial aid, there's – uh, that assist with the, um, full-time residents, and then sometimes they pay full. So you don't have the requirements met for this it kind of depends on what the population of the college is.

Overall, regardless of the justification offered, the interviewees understood that statistical hypothesis testing was not an appropriate means of answering the research questions stated in answer choices *b* and *c*. They did, however, differ in their decision to choose either *a* or *d*.

Most of the interviewees realized that *a* and *d* were very similar and admitted they struggled to make a decision between the two. In fact, Interviewees 29, 191, and 81 could give no real reason for choosing one over the other. Interviewees 81 and 191 both chose *a* and Interviewee 29 chose *d*.

Other interviewees could give a clear reason for making the choice they did. Five of the interviewees who chose answer choice *d* did so because there was a sense that answer choice *a* was too broad. Consider the following excerpts.

Excerpt: Interview 192
I: …Because *d* actually compares two specific things, to me.

M: Ok.

I: So, *a* is a bit more broad, a bit more general, a bit too general.

Excerpt: Interview 122
M: …So what made you chose *d* over *a*?

I: Because it would be, it's easier con – to conduct this, uh, statistic analyst because you have, uh, a population already regarding to the students. There are probably like 40 or 50 students in one class. So you'd be able to do more, a comparison between the two because…and it's less…and it's less time consuming to me also to try to make a hypothesis testing, because regarding to trying to walk, walk around or find students who wants to be…participate in the, um, this survey or the hypothesis test might take a long time.

Excerpt: Interview 15
M: Do you have any idea why people would pick *a* as opposed to *d*?

I: I guess you're still like comparing two different groups. But I feel like *d* is like a better choice because like it's a lot more specific like the groups are both learning the same things in class. The only difference is the times so you really only have like one variable. But with like an athlete and a random student, like the athlete could be going to games all the time and that's like why they have a lower GPA but the other student could be like a drug user who is just high all the time and doesn't go to class and they're gonna have a lower GPA. And, then like another random student could be like your top scholarship winner who is like going to every single class and doing awesome and … but like that doesn't mean that every other student is just like them. Like, there's just so many variations in the category of other students. I guess it's just not specific enough.

Excerpt: Interview 169
I: Ah, the reason I chose that one…ah….is because [M: You chose *d*?] it was a specific, and I chose *d*…it was the most specific of all the questions. And moreover *b*

> and *c* are not, to my knowledge something that can …you can test with hypothesis testing at least in STAT100. And, then *a* while along the same lines as *d*, isn't very specific…ah, which athletes, you know…you know in what courses…because there's, you know, there's a lot of variables to that equation that can't necessarily be answered for that, a statistical hypothesis test, versus *d* where it's…you're taking the same course. It's just at a different time. If you look at it from like a very scientific method point of view it's your in kind of variable is the time.
>
> **Excerpt: Interview 172**
> M: A lot of people struggled between *a* and *d*. In fact, very few people chose *b* or *c* [I: Yeah]. So it was almost half split, people who chose *a* and *d*. So can you talk about, is it, are you, is it possible to think about or talk about why you were more comfortable with *d* than *a*?
>
> I: Um…I guess…I was a little…I knew that these two, like the two sections that you were testing were kind of set numbers of people. Whereas, *a* was a larger population of people or, I mean, you could take a sample, I guess, from both but…I don't' know. I guess I found it like it's more…ah…sss….you're, I guess you could be more sure about the hypothesis test when you have like the exact same number of people…or about in the two statistics sections.

Each of these interviewees chose *d* because it was more specific in some sense. Interviewee 192 did not give a specific reason. She simply thought *a* was too broad. Interviewee 122 thought that the populations addressed be the research question in answer choice *a* were too large. She thought *d* was a better answer because it wouldn't be difficult to find students to participate. Interviewees 15 and 169 thought *a* was too broad in that there was too much variance in the populations described in the research question. Therefore, they thought *d* was the better option. Finally, Interviewee 172 was more comfortable with answer choice *d* because the researcher could be sure that the populations were approximately the same size. The populations addressed in answer choice *a* were too large.

With the exceptions of those who guessed, the interviewees who chose answer choice *d* over answer choice *a* did so because the populations addressed were smaller, more specific, and less variable. In contrast to this reasoning, two of the interviewees who chose option *a* (and who did not guess) did so *because* the populations addressed in that answer choice were larger. The following excerpts illustrate this reasoning.

Excerpt: Interview 112

I: Well, *b* - what equation predicts a student's freshman GPA from his or her SAT score – I pretty much like…hypothesis testing doesn't find equations. So I ruled that one out immediately. And, then *c* was for the typical costs and I thought that was more of an average type thing. So I crossed that one out. And, then, I was kind of divided on *a* or *d*, because it was both two different groups. One was athletes versus the other students and the other one was like…sect…noon section versus 2 pm section. I mainly just chose *a* because it seemed like an example from the textbook.

M: (laughs) Ok.

I: Yeah, cause it seemed like it was…like two larger groups and like the noon and 2 pm…that kind of just didn't seem like a real good reason to test it. I don't know why. It just kind of struck me that way.

M: Ok. Um…so mainly chose because it's an example from the text and it was two larger groups?

I: Yeah.

M: And, *d* seemed too small.

Excerpt: Interview 132

I: Um hum. (reads the problem out loud) Um…um…well I …I don't know why I decided between *a* and *d*. Probably because…um… *a* had a broader…ah…like population to go by because you're talking about all athletes versus non-athletic students. And, so you're hypothesis would be just what are the GPAs of all of the other students and then are athletes higher or lower than that.

Interviewee 112 was uncomfortable with answer choice *d* because the populations described were too small. He did not understand why someone would want to test it. Interviewee 132 also chose *a* because the populations addressed were larger than in answer choice *d*.

One other interviewee chose answer choice *a*, but did so for a different reason than that offered by Interviewees 112 and 132. Interviewee 77 initially chose option *d* but then changed her answer. The following excerpt illustrates her reasoning:

Excerpt: Interview 77

> M: Ok, so *a* you would have picked. [I: Um hum] You feel more comfortable with *a* now. [I: Um hum] You're allowed to change your answers, too, [I: Yeah] throughout. So why, now are you more comfortable with *a*?
>
> I: Because it's …you can make that direct claim out of this thing. Um…like you can say, you know, like with athletes and other students you can say…um like the alternative hypothesis would be mew1 less than mew2. So that's pretty…I mean that's a solid claim that you can make…and you could do a…hypothesis test…um…to find it out. This one it's…the wording is kind of vague where it says perform better…um…so, I guess, that's why I would switch to *a*.

As she re-read the item, Interviewee 77 had trouble remembering why she chose answer choice *d* over *a*. She ultimately decided to change her answer to option *a* because she did not know what was meant by "perform better" in answer choice *d*. So, because the measure of interest was clearly defined in answer choice *a* she chose that over option *d*.

Analysis

The explanations offered by the interviewees for their choices on this item confirm what was found in the quantitative phase. The interviewees understood statistical hypothesis testing to be a useful method to answer research questions that compare two groups. However, they did not understand the value of statistical hypothesis testing as an *inferential* method. It seems they were focused on minimizing the variance that might exist within populations in order to have a "more accurate" test. To do so, the populations should be smaller and more manageable. Unfortunately, the interviewees did not understand that statistical hypothesis testing is a powerful, inferential method that takes into account the variability that exists between and among samples making it most valuable when used to test hypotheses about large populations.

It should be noted that even when interviewees chose the correct option, their reason for doing so did not *necessarily* provide evidence that they understood the value of statistical hypothesis testing as an inferential method. Some of the interviewees guessed the correct answer and others merely said they thought it was better to use hypothesis testing to address research questions about large populations. In their explanations, however, they did not explain *why* this was the case. They did not mention the role of inference in statistical hypothesis testing. They did not mention that statistical hypothesis tests provide a means of dealing with the variability that exists between and among samples. Thus, on the basis of their response to this item, we do not know, for sure, whether these interviewees understood all of these ideas.

We now turn to another category of items classified in the ***context and method*** grouping: **Collect a Sample**. Analysis of student responses within this category might provide more insight into student understanding of sample variance and the role of statistical hypothesis testing in dealing with that variance.

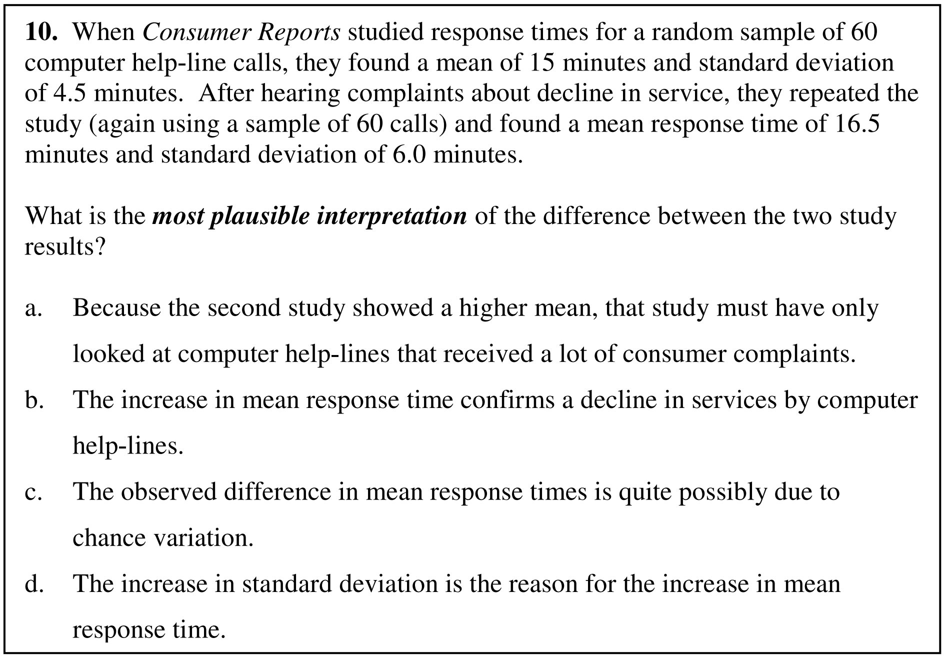
Collect a Sample

Items from the **Collect a Sample** category were written to assess whether introductory statistics students understand that samples are expected to vary and that, in order to apply statistical hypothesis testing, the samples must be randomly chosen and representative of the population. If the sample is randomly chosen and representative of the population, statistical hypothesis testing may then be used to determine whether the sample is unusual under the assumed null condition. The test, therefore, takes into consideration the variability that exists among samples.

Only one item from this category was included on the multiple-choice assessment. It was, therefore, also included in the follow-up interview. Item number 10 assesses whether introductory statistics students understand that samples vary and that it is, therefore, difficult to draw a conclusion about populations based solely on one sample. Statistical hypothesis testing is useful to draw an *inference* about a population based on information from a sample. However, any conclusions reached by a statistical hypothesis test are inferences, not confirmations or proofs.

Figure 5.8

Item Number 10, Multiple-Choice Assessment



10. When *Consumer Reports* studied response times for a random sample of 60 computer help-line calls, they found a mean of 15 minutes and standard deviation of 4.5 minutes. After hearing complaints about decline in service, they repeated the study (again using a sample of 60 calls) and found a mean response time of 16.5 minutes and standard deviation of 6.0 minutes.

What is the ***most plausible interpretation*** of the difference between the two study results?

a. Because the second study showed a higher mean, that study must have only looked at computer help-lines that received a lot of consumer complaints.

b. The increase in mean response time confirms a decline in services by computer help-lines.

c. The observed difference in mean response times is quite possibly due to chance variation.

d. The increase in standard deviation is the reason for the increase in mean response time.

The correct answer is option *c*. It reflects the notion that samples vary. Answer choices *a* and *b* are incorrect in that they attribute the difference in means to some characteristic of either the study or of the population itself. There is no reason to assume that the sample was not randomly chosen. Therefore, option *a* is not correct. Answer choice *b* makes a broad, general statement about the population based solely on information from the sample. This is not a reasonable interpretation. Answer choice *d* is incorrect in that the difference in means is attributed to the difference in standard deviation. The two are not related in this way.

Performance on this item in the multiple-choice assessment was relatively strong. Of the 104 participants, 10.6% chose answer choice *a*, 32.7% chose *b*, 45.2% chose *c*, and 11.5% chose *d*. These results indicated that a fair number of introductory statistics students understand that samples may vary. However, an almost equally large percentage of introductory statistics students think that one sample confirms a hypothesis. This issue was explored further in the follow-up interview.

Of the eleven interviewees, no one chose *a*, 4 people chose *b*, 6 people chose *c*, and 1 person chose *d*. Over the course of the interviews, only one person changed his answer. He changed from *c* (the correct answer) to *b* (an incorrect answer).

Interviewee Explanations: A Summary

The explanations offered by the interviewees for the elimination of answer choice *a* were very similar. By and large, they all felt that *a* was not a strong assumption to make based on the information that was given. Some examples of this reasoning are included in the following excerpts.

Excerpt: Interview 169
I: …Ok. Um…I still agree with my answer that…um…yeah…so I still agree with my answer…the more I think about it, the more I'm confused (both laugh). Um…um…*a*…um…if Consumer Reports…that just doesn't seem…that doesn't seem plausible. I mean, I, I…the question asks for the most plausible interpretation of the difference…um. [M: You don't think…] Because if you, if Consumer Reports, again, if they looked at a…um…if they looked at only computer help lines that received a lot of customer complaints it would, of course, be obvious that it would, that the mean and the standard deviation would go up. But, it's not a useful…it wouldn't have been a useful thing to do.

Excerpt: Interview 81
I: Um…yeah, I said *c*. I didn't think *a* was right it says because the second study showed a higher mean, that study must have only looked at computer help-lines that

received a lot of consumer complaints, cause it doesn't really say that. I don't think you can just assume that. I think that's assuming a lot.

Excerpt: Interview 191
I: …*a* says because the second study showed a higher mean, that study must have only looked at computer help lines that received a lot of consumer complaints. Um, I – I think you assume that they, um – they tested a random sample. Also, the fact that, um, although the, ah, mean response time was greater in the second sample, the standard deviation is greater, so that means that there were some times that, ah, were relatively low or significantly lower, um, so because of that, not all of them would have been, ah, lines that received computer or consumer complaints.

These examples are representative of some of the more common explanations offered by the interviewees. Like Interviewee 81, Interviewees 15, 77, and 122 didn't think it was a reasonable interpretation because nowhere in the description did it say anything to assume that the sample was not randomly collected. Like Interviewee 169, Interviewees 112 and 192 considered the context and didn't think that *Consumer Reports* would do things differently in the second study. Finally, like Interviewee 191, Interviewee 132 reasoned that, since the mean *and* standard deviation increased, it was possible that some of those help lines were better than the others. Therefore, the study did not simply consider those help lines for which there were complaints.

With the exception of 1 interviewee, none of the interviewees chose option *d*. Some of the interviewees reasoned that a change in standard deviation does not necessitate a change in the mean. This reasoning is illustrated in the following excerpts.

Excerpt: Interview 15
I: Um…they're trying to say that like…um…higher standard deviations means that you have a greater range in your data. Um...and so, like with a greater range you have more of higher numbers…and that could make your mean higher. But, like, some data sets…like if you have a very small data set…and like a certain mean…you could have one standard deviation that would be really small. But, if you have a large data set that's kind of like more distributed but still (inaudible) evenly then you could come up with the same mean but your standard deviation would be really big. So, like, just standard deviation alone can't really explain an increase or a decrease in the mean. Because it like…it goes both ways to the mean. Like, not just one way.

Excerpt: Interview 132

> I: …And, *d* doesn't sound reasonable enough. Just because your standard deviation gets larger, that means there's…the mean response time would change. Because, you don't necessarily have to have a lot of variation to increase your average. Did I answer it all clearly enough?

These excerpts indicate strong understandings of the definitions of and relationship between mean and standard deviation. This understanding was shared by Interviewees 122, 112, and 191 who reasoned similarly to Interviewees 15 and 132.

Other interviewees did not have such strong understandings of mean and standard deviation. Therefore, they did not choose option *d*. Interviewee 77 did not remember how the two were or were not related and, therefore, did not choose *d*. Interviewee 172 did not think that it mattered. Interviewee 81 thought *d* was plausible, but he liked answer choice *b* better. And, Interviewees 192 and 169 found the statement in *d* confusing because they didn't know how the two were related. Therefore, they did not choose *d* either.

Interviewee 29 did, however, choose option *d*. Her reasoning is illustrated in the following excerpt.

> **Excerpt: Interview 29**
> I: I think this was one of the ones that I kind of guessed on (laughs)….but…yeah (laughs). Um…I think I picked this one because…like there was an increase in standard deviation so when I think of standard deviation it's like how much…um…like data varies or deviates from a middle range, the average (laughs). So, I guess, think of that a lot more data deviated. It caused the mean to go up because…um…now because you have to, I guess, more data that was farther away from the average.
>
> M: Ok. Um…what about the other ones? Why didn't you pick the other ones?
>
> I: Um…I wasn't sure like…um…whether there was chance variation (laughs)…and…
>
> M: In general, or… or for this study…for this, these two studies? Like, I'm not sure, I just want to interpret what you said. [I: Right] You said you're not sure whether there was chance variation. Like, that it doesn't exist in life? Or, that it didn't exist for these two studies?
>
> I: I guess for these two studies because…I don't know. I mean you don't know all the factors that are involved in it…so. I guess cause I wasn't really sure if there was

> actually chance. Like, I mean, you know, it could be that, I don't know, the call…or calling this place like, you know, they just have bad service or something (laughs). You know, it's…so, yeah. That's why I didn't pick that one. I wasn't sure.
>
> M: Ok.
>
> I: And…(reads out loud, somewhat inaudibly)…and, then *b*…I felt like because you weren't sure of chance variation you're also…weren't sure of like factors involved like…so I mean…yeah they could have bad service too, but like you really don't know for sure. So it's hard to say like these were plausible interpretations.
>
> M: Ok. And then *a*?
>
> I: Also you don't know if they're consumer complaints (laughs). Yeah, I mean you don't know what they tested or you don't know the callers and you don't know whether it was because…of…um…consumer complaints.

This excerpt from Interviewee 29 is interesting in that it highlights the struggle she had with the notion that samples from the same population may vary by chance. She claims that she doesn't know all the factors involved and, thus, does not know if chance could be one of those factors. However, chance is *always* a factor and should always be a consideration. Statistical hypothesis testing is useful in attempting to quantify the degree to which the variation may be due to chance, but even it doesn't eliminate the possibility.

The issue of "chance" was problematic for other interviewees and led to the elimination of answer choice *c* in favor of answer choice *b*. Interviewee 112 did not think that answer choice *c* should be considered as an option.

> **Excerpt: Interview 112**
>
> M: And, *c*…
>
> I: I didn't think you should just assume it's due to chance variation.
>
> M: Why not?
>
> I: I wasn't really sure (laughs). It just didn't seem right to assume anything

In contrast to Interviewee 112, Interviewees 169, 81, 191, and 172 all recognized that *c* was a possible answer. However they felt that the difference in mean times justified their choice of answer choice *b*. The following excerpts illustrate this thinking.

Interviewee 169
I: Um…I don't…quite understand *d* but it doesn't seem right to me. And, then, let's see…um…and *c* seems kind of like a cop, a cop out to me. It seems like, yeah, I mean, it could be it…but based on the fact that…and it's asking due to chance variation…and then…but it doesn't seem like that would be necessarily something that would raise it that much. It's a pretty decent increase in mean time…doesn't seem right to me.

M: So you think the…the fact that it's so big…*b* confirms it, that…?

Interviewee 81
I: …Um…let me see…mmm….huh…actually it'd probably be *b*. Um…god, I misread that one, yet again. Way to go. Um…I would definitely say *b* or *c* and I'm leaning closer towards *b* now.

M: Ok. Why?

I: Um…because it says increasing mean response time confirms a decline in services by computer help-lines. Um…I mean, the fact that it did…you have to wait an extra minute and a half after everybody complained about it. Ah…shows that there is…has been a decline in services. And since the, ah, standard variation went up 2, it sort of shows a decline because it means, you know some people are, you know, might be…helping you earlier but some people might be helping you even later.

discussion continues

I: Um…chance variation…I think it's cause I misread *b* as an answer. Um…if I…had read it properly I definitely, I definitely would have chosen *b*. Um, I just…I just said chance variation cause I mean, it could be per chance, it could just happen to be that study was different. No, really, you know, you do multiple studies of stuff like this. It's not always gonna be the same thing.

Excerpt: Interview 191
I: – um, of calls, so the sample sizes are the same. Um, and they found that the second study, there was, um, a larger mean response time. Um, which would suggest that there's a decline in service, um, so I think that's why I chose *b*

discussion continues

I: *c*, the observed difference in mean times is quite possibly– *c* is possible, but, um, again, the chances of that, um, are low.

Excerpt: Interview 172
I: …And, um…I chose…I mean, I guess *c* is plausible too because it just says it's due to a chance variation so, I, I mean…it could be because it is just a sample. I mean…it, it could increase or decrease just based on chance, like it says, you know, depending on the people that they get. Um…but I think I chose *b* just because…it's talking about a…um…response times. So, I was thinking of it like…they didn't respond like the mean time before was 15 minutes they hadn't responded…and then the next mean was 16 minutes. So, I was thinking of it as…um…they…put them on hold, I guess, like even longer and…so…the increase in the mean response time confirmed a decline in, like, the

> services…so. I mean, if they're keeping you on hold longer then…it's a decline in service so…I guess that's how I thought about it.

Each of Interviewees 169, 81, 191, and 172 thought that *c* was a *possible* answer. However, they thought that since the sample data supported the statement in answer choice *b*, it was the best answer. Again, we see some discomfort with the idea that the difference could be due to chance. Interviewee 169 thinks that it is a "cop out" to make that claim and Interviewee 191 thinks "the chances of that are low". Unfortunately, this claim is not a "cop out" so to speak and should not be discounted, even if a more formal comparison of the samples were made through a statistical hypothesis test.

Five interviewees chose the correct answer, *c*. However, 2 of those interviewees also thought that answer choice *b* was correct. Excerpts from these interviews are included here.

> **Excerpt: Interview 77**
> I: Um…I said *c* because…um…ok. I said *c* because it's possible that, you know, there might have been…there might have been a longer time for the calls to be responded to…depending on the length of calls…you know, the length of each call and…how many people were working the telephones. Um…so I definitely think that there is a chance due to variation. Um…it's possible that it could have been *b* also because it does say that…in the second sample…the mean time did prove to be higher than the first sample, which could show that there's a decline in service.
>
> **Excerpt: Interview 132**
> I: Ok. (reads out loud, somewhat inaudibly) Ok…I said *c* the observed difference…um…make…um…*a* seemed,…ah…like the answer just seemed…seemed complicated so…ah…because ah…yeah…so that's why I didn't choose *a*. Ah…the increase in time…*b* is saying that…that like…they've gotten worse at receiving complaints, ah…which is true. Like they're…they're…ah…response time got longer, but the variance was also much higher…the standard deviation was 6 minutes. And, that's why I chose *c* because that means that there was a lot of calls that were very short…were shorter than how they were originally. But, they did have some longer ones. So it could have been just…ah…problems with those specific 60 calls that they sampled from.

Both Interviewees 77 and 132 thought that the increase in mean times confirmed a decline in service. However they each thought that *c* was a better answer. Interviewee 77 did not clarify why this was the case for her. Interviewee 132, on the other hand, considered the change of standard deviation to justify her thinking that the difference in means may have just been due to the sample.

Interviewee 122 also thought that answer choices *b* and *c* were correct. However, her reason for choosing *c* is due to misinterpretation of the phrase “chance variation”.

> **Excerpt: Interview 122**
> I: Because some calls can last for three minutes, one, sometimes could last six minutes, so it just depends how long, um, the calls are. So that couldn’t be a. Um, b, the increase in mean response time confirms the decline in services by computer help lines. I think b makes sense also. Um, b, makes sense. The increase in mean response time confirms the decline in services by computer help lines because every time like, in for a call center, or, um, help line, like if you have a longer mean, then…you’re supposed to have the shortest amount of time too, but use it effectively, ‘cause if you have a long phone calls that you’re kind of like, um, like being more effective because the increase in mean offers a decline services. Yeah, but you’re not being able to get more calls in also. Um, c, the observed difference in mean response time is quite possible due to chance variation, and the variation is in regarding to the standard deviation, and, um, standard deviation of the first sample of 60 is 4.5 minutes and the second standard deviation is 6 minutes. Most likely, like, with my, like, trying put, like cons – like my experience or put some constant into it, like, um, the best work help line it’s best to have the, um, the least the shortest amount of time. Like it’s a big, huge jump from 4.5 to 6 minutes. It makes a big difference regarding to the variation between the two.

Interviewee 122 interpreted the word “variation” to mean variance, or standard deviation. She thought that answer choice b was correct but felt that answer choice c was a better answer. In her experience working help lines, she knew that a change of standard deviation from 4.5 to 6 minutes was a big difference. Therefore, she chose option c. This response is interesting in that Interviewee 122 drew on her own experience to answer the question. However, she misinterpreted the meaning of answer choice c.

Two interviewees chose answer choice c and did not think that answer choice b was acceptable. Their reasoning is illustrated in the following excerpts:

> **Excerpt: Interview 192**
> I: Ten. (Reads number 10 out loud). I chose *c*. I think I was thinking along the lines of chance variation. Something could have happened there…could have

been…something happened that day…or whatever that caused the variation in…in the times.

discussion continues

I: Um…*b* was…I thought was ridiculous, so….

M: Why?

I: I mean…just because it takes…to me when it, decline in service by computer help lines…that like…they're helping you with situations, not the time that it takes for them to help in the situations, because you're still being helped. So it doesn't affect quality of help therefore it doesn't affect the service, in my opinion.

Excerpt: Interview 15

I: Um, and then like *b* you can't really confirm anything because you're just comparing them and they obviously have different standard deviations. So, like it's not necessarily true. So you can't really confirm it unless you do a test about it. …And like with *c,* like chance variation happens because you're just taking a sample. You're not using all of them. So, like…it's a possibility. And without looking at like anything else, just looking at like here are the two means from the two samples with different standard deviations…like, that's something that you could say…and like I think that that's something you would say and people would agree with you. They'd be like yeah that could be true.

In eliminating answer choice *b*, Interviewee 192 considered the context and claimed that longer phone calls may or may not have affected the quality of the service. Therefore, she did not think that a difference in mean times implied a decline in service. Interviewee 15, on the other hand, eliminated answer choice *b* because such a broad claim could not be made using only a sample of the population. Regardless of their reason for eliminating answer choice *b*, Interviewees 192 and 15 both recognized that the sampling process could produce samples with different means by chance. Thus, they chose answer choice *c*.

Analysis

In summary, the explanations offered by the interviewees for their choices on Item number 10 indicated that, though the interviewees recognize that samples may vary, they tend to attribute the difference in samples to a difference in the populations. Unfortunately, many of the interviewees were uncomfortable attributing a difference in sample statistics to chance. They were willing to draw conclusions about a population based solely on information from a sample. Such a lack of understanding could impact the way in which they interpret the results of a statistical hypothesis test and, ultimately, use those interpretations to make decisions in the real world. Analysis of student reasoning in the final category of this section provides more insight into the ways that students interpret results of a statistical hypothesis test in context.

Implication for Practice

Items from the **Implication for Practice** category were written to assess whether students understand that a statistically significant result does not necessarily imply practical significance. Before making real world decisions, various factors should be considered besides the fact that a hypothesis test found a result to be statistically significant. Type I and Type II errors, effect size, and sample size should all be considered when taking action based on the results of a hypothesis test.

In the quantitative phase, performance on items from this category was fairly weak. Two items from the **Implication for Practice** category were included on the multiple-choice assessment. However, it was hypothesized that students misread one of the two items and chose answers that were incorrect for the wrong reason. Nevertheless, results from the quantitative phase indicated that introductory statistics students do not have a strong understanding of the factors that should be considered when interpreting the results of a statistical hypothesis test in the "real world". It was hypothesized that because students do not understand the role of probability and inference in statistical hypothesis testing, they have difficulty considering the role of context when using the results of a test to make decisions in real world context.

In the follow-up interview these issues were addressed to gain more insight into student thinking. One item from this category was included in the follow-up interview: Item number 12. Item number 12 was written to assess student understanding of the factors to consider when using the results of a statistical hypothesis test to take action in the "real world".

Figure 5.9

Item Number 12, Multiple-Choice Assessment

12. In an educational study the mean test score of students studying from a new, experimental textbook A was greater than that for students studying from a previously used, traditional textbook B, with significance at the $\alpha = 0.05$ level.

What action in response to that result ***makes most sense*** to you?

a. Compare the mean scores to see if the difference is great enough to merit the cost of new books.

b. Schools should adopt textbook A because its use leads to significantly better learning.

c. Re-analyze the data to see if the difference in means is significant at the 0.10 level.

d. Take no action until the study is repeated, because the difference in scores could be due to chance.

The correct answer is answer choice *a*. This option highlights the fact that statistical significance does not necessarily imply practical significance. One must take into consideration the effect size by analyzing the actual difference in means with respect to the sample size. Answer choice *b* was written to assess whether students think that action should be taken on the basis of the results of a statistical hypothesis test, without further consideration of other factors. Answer choice *c* was written to identify those students who do not understand the relationship of the decision rule to the conclusion of a statistical hypothesis test. Answer choice *d* was written to assess whether students understand that, in reality, a study such as the one described is costly and time-consuming. Statistical hypothesis testing is a powerful tool for comparing two groups based on analysis of *one* sample. With such a tool, it is not efficient to conduct a costly, time-consuming study again.

In the multiple-choice assessment, 36.5% of the participants chose answer choice *a*, 20.2% chose *b*, 19.2% chose *c*, and 24% chose *d*. Hence, the participants chose a variety of answers, very few of which were correct. These results indicate a range of potential ways of reasoning about this item and the follow-up interviews were very helpful in uncovering these ways of reasoning.

Of the eleven interviewees, 6 chose *a*, 2 chose *b*, 1 chose *c*, and 2 chose *d*. Unfortunately, over the course of the interview, 1 person changed his answer from *a* to *b* and one from *a* to *d*. Ultimately, 4 people had chosen *a*, 3 people had chosen *b*, 1 person had chosen *c*, and 3 people had chosen *d*.

Interviewee Explanations: A Summary

There seemed to be a general feeling among those who chose *a* or *d* that something should be reanalyzed. Some understood that the suggestion in answer choice *c* would not be helpful in that the difference would still be significant at a significance level of 0.05. Some did not. Generally, though, the interviewees thought some kind of analysis was appropriate before deciding to buy the books (as suggested by answer choice *b*).

Of the 4 interviewees who settled on answer choice *a*, 2 eliminated *c* because it would result in the same conclusion, 1 wasn't sure what *c* meant, and 1 thought *c* was a reasonable, but not the best, option. Various justifications were given for eliminating answer choice *b*. Most of the interviewees thought that answer choice *d* was reasonable but thought that *a* was the best option. Examples of the explanations given for choosing answer choice *a* are given in the following excerpts.

Excerpt: Interview 132
I: (Reads 12 out loud, somewhat inaudibly) Um…I chose *a*…ah…because it seemed…well it made the most…made the most sense…ah. I don't know, because if the mean scores are better with the new textbook that means it's a better instrument for education. Ah…*b* says the same thing but it doesn't really say it when you're talking about data. It's just saying it's better for learning. So it doesn't really directly talk about data…or…calculations. I almost chose *c*…that is significant…um…so, if I…if I made my acceptance region smaller that means that'd be more chance I could accept my…um…alternate hypothesis and so that it'd seem reason…reasonable.

M: Ok, so if you made which region smaller are you talking about?

I: My acceptance region for my null hypothesis. That means I make my acceptance for my alternate hypothesis larger and so I would be wanting to make my acceptance region for alternate hypothesis smaller to see if it would still work. And, *d* just, I don't know, seemed silly. I don't know. Take no action until the study is repeated. Um…well they didn't really talk about variance in the first place. Well…I mean…(inaudible).

Excerpt: Interview 15
I: Um…I guess you're using…cause like with the little description before the question…the purpose of this study is to see if you wanna get these new books or not. So, if you did this study and found out that the new books are better than the old books and there is a significance to it, then um…like…practically you'd wanna…like take no action until the study is repeated is just kind of silly cause what if it's never repeated and you're just sitting around forever and nothing's ever going to get done.

M: (laughs) Ok

I: And, then, um…like…I guess like you could sit there and reanalyze it but…like…if it's like …so what if it's not…and so does that mean like…oh because it's not significant at our chosen level….cause the levels are kind of a little bit arbitrary like different people have different preferences for levels. So, like…just because it may not be significant, if you test it again at the 0.1 level doesn't mean that there's some gain that you could get out of these books….I guess like especially in a school where budgets are kind of tight. If you do like a cost analysis and see like oh these books are only a dollar more than the old books, then they're probably worth getting. If they're like a thousand dollars more then the old books, then you may not want them and then…like…if you do that, then you may want to reanalyze it and see like oh well it is greater at the 0.1 level, too, like maybe you can better justify spending a greater amount of money on them.

Excerpt: Interview 81

I: All right. Um, I said *a* compare the mean scores to see if the difference is great enough to merit the cost of new books. Um…I basically said, well the, the experiment is…is that textbook A was greater than…well…all right, sorry. In an educational study the mean test score of students studying from a new, experimental textbook A was greater than that for students studying from a previously used, textbook B, with significance at…alpha equals, ah, 0.05 level. And, what action in response to that result makes most sense to you. Compare the mean scores to see if the difference is great enough to merit the cost of new text books. That's sort of what they did…um… and *b* says schools should adopt textbook A because its use leads to significantly better learning. *C* says re-analyze the data to see if the difference in means is significant at the 0.10 level. And, *d* says take no action until the study is repeated, because the difference in scores could be due to chance. Um…I guess *a* made most sense cause you could….you could compare them enough to see how much greater it is, if it's actually worth getting new books. I think that's why I chose *a*…cause it seemed…cause, like, sort of because getting new books seems like what somebody would do after this experiment, which is why I chose *a*.

M: Ok. And, why not *b*, *c*, or *d*?

I: Um…because…it didn't really say how, if it was significant enough…to…adopt textbook A…for better learning. And…for *c* it says…um…re-analyze at the 0.10 level and, of course I really, as in question 11, I really wasn't sure how that would make a difference so I didn't choose that answer.

Excerpt: Interview 122

I: Okay. Twelve is, um, is asking what makes the most sense regarding to in the education study the mean test scores of students studying from a new experimental test book A was greater than that for students studying from previous used, um, traditional text book B with significance of – off of 0.05. I chose *a*, compare the mean scores to see the difference great enough to merit the cost of new books. And I chose that because when you look at the other ones, like B, schools should adopt textbook A because it use – it's use leads to significantly better learning. Well, they can't really actually find, um, because they don't actually have the exact numbers to see if it's great – actually how much greater it is by using textbook A and textbook B.

M: Okay.

I: In my opinion. On *c*, reanalyze the data to see if the first mean is significant at the 0.10 level, well if you're trying to reanalyze that actually will cover some portions of the 0.05 'cause that's 95% of that, um, the results. So that will still be included there.

M: Okay.

> I: Um, *d*, take no actions to the studies repeated because of difference in scores should, um, scores could be due to chance. Um, well, it may be possible because it is a random sample, but regarding to the student test score what are they studying.

These excerpts highlight the fact that though the interviewees all settled on answer choice *a*, they had different reasons for eliminating the other answer choices. Interviewees 81 and 122 eliminated *b* because they weren't told the exact numbers so they didn't know whether they could make the claim in statement *b*. Interviewee 132 eliminated *b* as opposed to *a* because *a* mentioned "mean scores" whereas *b* did not. None of the interviewees eliminated *b* for the correct reason. It should have been eliminated because it was too broad a statement. Additionally, answer choice *d* was eliminated for the correct reason by only one interviewee, Interviewee 15.

Of the 3 interviewees who chose answer choice *d*, only Interviewee 77 eliminated *c* for the correct reason. Two interviewees, Interviewees 77 and 192 eliminated *b* for the correct reason. They all thought that *a* was reasonable, but that *d* was the best option.

> **Excerpt: Interview 77**
> I: Ok…um…it's definitely not *b* because this was just from one study…it's not like it was a repeated study of the compare, the comparison of the two textbooks and the mean test scores. Um…and I didn't think it was *c* because if you use…the significance level of alpha being 0.10…um…then you have a greater chance that you will incorrectly reject the null hypothesis…um. You should actually, if you want to reanalyze the data you should test it with an alpha that's smaller, not greater.
>
> *discussion continues*
>
> I: Um…I guess because it was just…so far it seemed like it's been only one study…and…I guess when you, when I'm looking at *a* and *d*, since *a* is talking about the cost of new books and, you know, trading the newer ones for the traditional ones were…gonna cost more…if you were going to do that you, it would make sense to repeat the study to…be absolutely sure that there is a difference in the mean scores and that it wasn't just due to chance. Um…and once you find that, like if….you did…more studies and found that…the difference in the, there was a significant difference in the mean scores than that would…that would kind of…I guess, give you enough reason to switch out the textbooks.
>
> **Excerpt: Interview 192**
> I: (Reads number 12 out loud) Ok…um…I chose *d*, take no action until the study is repeated, because the difference in scores could be due to chance. Um…(reads out

loud, inaudibly)…ah…I think I got my answer…compare the mean scores to see if the difference is great enough to merit the cost of new textbooks…I didn't choose that one, because I think they can't just do a test one time and say that it's correct. And, *a* kind of is just going off of those answers off that one test. *B*, schools should adopt text A because it leads to significantly better learning. Again, you can't act on it unless you test it more than one time. *C*, reanalyze the data to see if the difference in means is significant at the 0.10 level…yeah, you could do that but I would prefer to retest it again just to make sure that you get the same answer. [M: Ok] So I stuck with *d*. So you would repeat the study.

Excerpt: Interview 29
I: Um…I really, I picked *d* because…I mean, I don't know how they actually use test hypoth…I mean, yeah…I don't know how they actually use test statistics and what they do when, like they find one to support…I mean one who's gonna support it or reject it. So I guess I was thinking because, you know, there's still like you can find a conclusion but it could, you know, still be wrong. So, I guess you need to verify it or you need to validate it so…I guess that's why I picked *d* because, you know…to repeat the study…you can see if it's still supported and…I guess could help confirm whether your original conclusion is still…you know…right.

discussion continues

M: So why didn't you like *a* and *b*?

I: Cause I don't know how you compare…like…like how would you….decide whether like the difference is big enough to…buy new books? Like, I wasn't sure like how you would decide that…so I didn't…I didn't pick that one. And…ah…and, then, *b* kind of goes with it about…like how do you know…like how do…like how do you know the difference is big enough so that you should get brand new books or something.

As illustrated by the excerpts above, the three interviewees who chose answer choice *d* had different reasons for eliminating *a*, *b*, and *c*. However, they all thought the study should be repeated.

Three interviewees chose answer choice *b*. Among the three there was a general consensus that a significance level of 0.05 was low enough to take action. Two of the three interviewees eliminated answer choice *c* for the correct reason. They each gave different reasons for eliminating *a* and *d*. Their explanations are included below.

Excerpt: Interview 112
I: (reads out loud, somewhat inaudible) Well you don't…you don't know anything about the cost of the book. Like textbook A could be cheaper for all we know so that's why I didn't…disregarded *a*. (reads out loud, somewhat inaudible) And you already know it's significant at the 0.10 level if it's significant at the 0.05 level. That's why you

can x out *c*. And then it is significant…statistically significant at 0.05 so I didn't think you had to repeat the study…I chose *b* because it is…it does lead to significantly better learning…yeah, based on the study.

M: Ok, let me make sure…you went through those kind of quickly. I wanna make sure I got them…so you're disregarding *a*.

I: Yeah, cause we don't know what the cost of these books are.

M: If you knew them?

I: Yeah maybe if you knew A was like, incredibly expensive then…but then that would bring in *d* also. You'd have to redo it and maybe make it like 0.01. But, for all we know A could be less expensive.

Excerpt: Interview 191

I: Um, in an educational study, the mean test score of students studying from a new experimental textbook, A was greater than that for students studying from previously used textbook, B…significance at alpha .05. Response, the reaction in response to that result makes most sense to you. Ok, so without reading these options, I would say, ah, you can be confident because it's at .05, it's a low alpha level, so that means there's a high confidence level. Um, you can be confident, um, that, um, textbook A is better than textbook B, um, so I chose *a*, compare the mean scores to see if the difference is great enough to merit the cost of new books.

Um,…*b* schools should adopt... Well, I guess I think *a* would be wrong, because that's what the test – that's what this information is a result of. Um, the – the question already tells you that they already did a test and that the test scores for A were better than that of B.

M: Ok.

I: So *a* has already been done. Um, ok, I wouldn't do *c* because, ah, that means there's – you're less confident, so if alpha's at .05, you're 95% confident. If it's at .1, you're 90% confident.

M: Ok.

I: *b* and *d*, um, I would do *b*, because you can be pretty confident, um, in the information. *d* I would do if – if, ah, it was done at a higher, um, alpha level.

Excerpt: Interview 169

I: Hmm. Now I'm not quite sure of my answer…um, mainly because it doesn't…um…it doesn't say, the problem does not mention how much better in any way shape or form…and therefore, it's kind of difficult to…talk about, um, action or…the, the difference in scores, things about that. Um…I think that if, like, based on like…the circumstances that this italic section up here mentions that…um…teachers and school administrators are always interested in helping out their students and increasing their success…then, I guess that's why I chose *b*…just based on the circumstances behind the problem. But, if, if I'm looking at it from a purely analytic method

> then…then…um…*a*…*a*, *c*, and *d* all seem plausible…like they're very, very plausible…I mean, and like ah hypothesis testing, like….
>
> M: But you would want to…you think that *b* is the best option…just go ahead and buy the books because the test showed that it was significant at the 0.05 level.
>
> I: Yeah, based on the circumstances that it mentions up at the top. I put myself in their shoes, you know, if I have the money, if I have the, the means, then…go ahead…give it a shot, I feel like embracing technology and things like that, you know…um…they're always good things to do. I feel like a lot of public schools don't do that, having spent 12 years in the system…13, I guess, but.

These excerpts highlight the different justifications used by Interviewees 112, 191, and 169 for eliminating answer choices *a* and *d* in favor of answer choice *b*. Interviewee 112 eliminated *a* because he didn't actually know the cost of the books. If he knew the cost, and they were expensive, then he would want to take the action suggested in option *d*. Otherwise, in choosing one answer, he thought that *b* was reasonable because the conclusion to the statistical hypothesis test was that the books were different. Interviewee 191 thought that if a significant difference was found at a significance level of $\alpha = 0.05$, he could be confident in the results and take action. He thought this level of significance was low enough. Answer choice *a* was eliminated because he interpreted that to be the same as *b* and answer choice *d* would be reasonable if only the test were done at a lower (he said higher – but meant lower) level of significance. Finally, like Interviewee 112, Interviewee 169 eliminated answer choice *a* because he didn't know the cost of the books. He thought answer choices *c* and *d* were both reasonable. However, based on his experience in the school system, he thought the school should simply buy the books, if they have the means.

One Interviewee chose answer choice *c*. However, in her explanation, she struggled to justify her choice. The following excerpt illustrates her reasoning.

Excerpt: Interview 172

> I: Ok. Um…ok, um…well I thought that they were all…or…except for *b*, I mean because *b* I think you should, at least in this case, reanalyze it. Um…and the other ones…are still like talking about thinking about something else before just like buying, you know, the new text, textbook. Um…I think I just came down to *c* and *d* because I thought it, the thing that made sense to me the most was to repeat it…to you know, to just make sure that it was the same and *d* made sense because…ah…the difference in scores could be due to chance, like…there's always like an…ah…element of chance, I guess…and then it based on the, ah, the test scores of certain students if tested maybe a different class it could come out differently. But, um…I think I chose *c* just because it was…um…a different level of significance so…you could see if, um…like…I don't know…the way I have to think about this on the curve…um…(drawing on paper) so then like this in here is like 0.05 and I…at 0.05 it was rejected…so…then it here is…0.1…and…I don't know, I would have to like see, cause if it's…if whatever the test statistic like turned out to be, like it doesn't really tell you here, um…was like close to being accepted or rejected based on the like 0.05, you know…maybe like here's the rejection region. If it was like right here, I mean if it was like right here or something…maybe if this…I don't know. I was thinking just based on, like, if it was just outside it or…just inside it…or something, I don't know, that based on um…if you had it at a different significance level, like…if was inside it here and it was in, still in the rejection region in this, then you knew that…um…it wasn't just like…I guess…right on the boarder of being accepted or rejected. I don't know I was just thinking, I guess, in terms of…you test within a greater um…um…interval, I guess, to see whether it definitely should be rejected or maybe it could be accepted.

Interviewee 172 thought that more analysis should go into the situation than answer choice *b* would suggest. Therefore, she thought that *a*, *c*, and *d* made sense as next steps. In particular, she thought that *c* and *d* were the best options. As she tried to justify her choice of *c* over *d*, she seemed to be confused. She was not able to justify her reasoning. Unfortunately, at this point, the hour was coming to an end and there was not time to ask more questions to probe her thinking.

Analysis

The explanations that interviewees gave for this item were interesting in that there was a great deal of variation in their reasoning. Answer choice *c* was clearly an incorrect answer because it would not be helpful. The difference was found to be statistically significant at $\alpha = 0.05$ and, therefore, would be found to be statistically significant at $\alpha = 0.10$. However, though the vast majority of interviewees did not choose this option, many interviewees thought that answer choice *c* was reasonable. This result is surprising in that many of them answered item number 11 correctly and seemingly understood this phenomenon, on some level. Admittedly, the discussions surrounding item number 12 were not as long as others due to the fact that the hour was coming to an end. Given more time, the interviewees might have recognized their error.

Answer choice *b* was written to be a clearly incorrect answer. It makes a broad claim about a population based solely on the results of a statistical hypothesis test. These tests are not proofs. There is room for error. This should be taken into consideration. If, as in other items on the test, the interviewees did not interpret the statement in answer choice *b* to be a broad claim resulting from a notion of statistical hypothesis testing as proof, this option is not clearly incorrect. And, in fact, three of the interviewees chose this option. It does not necessarily mean that these individuals assume that statistical hypothesis tests prove a claim. They thought the other answer choices were reasonable, given they were provided with the necessary information and/or resources.

The choice between answer choices *a* and *d* was not so clear. And, in fact it was in deciding between these two options that many interviewees struggled. There was a general feeling that action should not be taken based on the results of one study alone. More analysis should be done. However, answer choice *a* is, in reality, the only *reasonable* choice. These studies are costly and time consuming. In addition, no matter how many studies are done, there will always be a degree of uncertainty associated with the results. Therefore, answer choice *d* is not the *best* answer. Though a couple of interviewees recognized this to be the case, many did not. It seems introductory statistics students do not have a good handle on the factors involved in conducting and interpreting results in a real world setting.

Summary

Overall, the data collected in the qualitative phase of the study confirm findings in the quantitative phase. Introductory statistics students do not have a strong understanding of the role of inference in statistical hypothesis testing and, therefore, do not fully understand the value of statistical hypothesis testing as an *inferential* method. They are uncomfortable reasoning about populations based solely on information from a sample. In addition, introductory statistics students do not have strong understandings of the factors that must be considered in making real world decisions based on the results from a statistical hypothesis test.

The qualitative phase, however, provided more insight into student reasoning about the role of context in statistical hypothesis testing. The real world is "messy" so to speak. There is great deal of variability and uncertainty associated with real world data. Statistical hypothesis testing is a valuable tool used to get a handle on that variability and uncertainty. The method helps researchers make a conclusion about the population through consideration of the fact that samples do vary. Through the follow-up interviews we see that students really do not have an understanding of these ideas. Though, on some level, they understand that sample data is variable, they are very uncomfortable with this idea. As indicated by the discussion surrounding item number 10, introductory statistics students are hesitant to attribute a difference in sample data to chance, whereas the assumption that the difference in sample data is due to chance should be the *first* assumption one would make. Formal statistical analyses, like hypothesis tests, are then used to determine to what degree “chance” can be ruled out as an explanation.

Without such an understanding of the value and power of statistical hypothesis tests as a means of making conclusions under conditions of high variability and uncertainty, introductory students do not *fully* understand the relationship of the context and the method. This lack of understanding impacts the way they interpret and take action based on the results of a statistical hypothesis test in a real world context. And, it influences their understanding of when statistical hypothesis testing is an appropriate method to use in research. The evidence provided by both the quantitative and qualitative phases confirms this notion.

Research Sub-Question Number 1

Results from the quantitative phase indicated that though introductory statistics students are able to perform the procedures associated with statistical hypothesis testing, they do not necessarily have strong understandings either of the logic and concepts that support those procedures or of the uses of the method in real world contexts. Students generally scored well on the course exam that assessed (mainly) student ability to apply the procedures to well defined, traditional statistical hypothesis testing problems. They did not, however, score well on the multiple-choice instrument that assessed student understanding of the overall logic, associated concepts, and uses of statistical hypothesis testing. In addition, the correlation coefficient between the two sets of scores was low. Hence, it was concluded that students who can perform the procedures do not necessarily have strong understandings of the logic, concepts, and uses of statistical hypothesis testing, and vice-versa.

However, as was demonstrated in the previous sections, scores on assessments tell only part of the story. More information on the *nature* of the relationship between student understanding of the procedures and of the logic, concepts, and uses of statistical hypothesis testing can be obtained through analysis of the follow-up interview data collected in the qualitative phase of the study.

The data used to address the second and third research sub-questions presented in the previous two sections can be used to address research sub-question number 1. In their explanations of their reasoning for the multiple-choice items, the interviewees often referred to procedures to justify their choices. This data, therefore, provides information on the relationship of the interviewees' understandings of the procedures and of the logic, concepts, and uses of statistical hypothesis testing.

Understanding of the Procedures and of the Logic and Concepts

Analysis of the data associated with the ***logic and reasoning*** items indicated that the interviewees did not have strong, connected understandings of the roles of indirect reasoning, probability, and inference in statistical hypothesis testing. When they were asked to explain the overall logic and reasoning of statistical hypothesis testing, the interviewees were not able to give explanations that demonstrated deep understanding of either the logic of hypothesis testing or of the concepts and reasoning that supported that logic. Instead, the interviewees often relied on their memory of procedures to explain statistical hypothesis testing. It was as if the interviewees used knowledge of procedures to "fill in the gaps", so to speak, of their knowledge. Unfortunately, they were not able to go beyond the procedures and give explanations that evidenced a deep, connected understanding of the logic and concepts that supported those procedures.

One of the points at which interviewees relied on procedural understandings in the interviews was in their attempt to explain why the null hypothesis should be a statement of equality and the alternative a statement of inequality. We saw in the discussion of research sub-question number 2 that the interviewees often cited rules for establishing the null and alternative hypotheses. They claimed that the null hypothesis always included an "equal sign" and represented the "status quo", so to speak. The alternative hypothesis, on the other hand, represented the claim the researcher is trying to demonstrate. Thus, it represents a change, and should be a statement of inequality. When asked whether the two could be switched, many of the interviewees thought that they could. This thinking provides evidence that, though the students know how to set up the hypotheses, they do not know how that "procedure" connects with the overall logic and reasoning associated with statistical hypothesis testing. The hypotheses are arranged in that way so that indirect reasoning may be used. The null hypothesis is clearly defined with a statement of equality. Hence, the assumed population is clearly defined and sampling distributions can be used to determine whether or not the sample presents an unusual occurrence. The interviewees do not seem to have an understanding of these concepts and ideas.

Another point at which interviewees relied on procedures to justify their thinking was in their explanation of the reasoning that supports the transition from the analysis of the sample to the statement of the conclusion. A great deal of time was spent with the interviewees in discussing this piece of the logic of statistical hypothesis testing. As the interviewees explained their reasoning, they referred to rejection regions, *z*- and *t*-tables, and significance levels. However, as indicated by the data presented in the discussion surrounding research sub-question number 2, the interviewees struggled to explain what these concepts were and/or why they were used.

When asked about rejection regions, the interviewees typically recited a set of rules used to determine whether or not to reject the null hypothesis. These rules inevitably referred to statistical tables and levels of significance. When asked what these concepts were, the interviewees either said that they didn't really know (they simply used them) or they recited more procedures to explain how they are used. In reciting more procedures some of the interviewees drew graphs, some sketched the outline of the tables, and some performed calculations. Using these tools, the interviewees explained how to use the level of significance, α, to decide whether or not the null hypothesis should be rejected.

The word "probability" was curiously absent in many of the interviewees' explanations of rejections regions, z- and t-tables, and significance levels. Most of the interviewees talked about the degree to which one is confident, or certain, of his final conclusion. They linked what they referred to as the "degree of confidence (or certainty)" to the significance level of a given statistical hypothesis test. The interviewees explained that the statistical tables were used to determine whether to reject the null hypothesis or not and the level of significance was used to indicate the degree to which the researcher is certain in his/her decision. Only a few interviewees mentioned the term "probability". And, of those interviewees, none of them indicated that the probabilities in the table were conditional probabilities, conditioned on the null hypothesis. Overall, none of the interviewees explained that probability was used to determine whether the sample was unusual, conditioned on the null hypothesis. However, most of the interviewees could recite the steps required to make a decision about the null hypothesis.

Another term that was absent in the interviewees' explanations was that of sampling distributions. None of the interviewees referred to sampling distributions in their reasoning. As we saw in the data associated with item number 9, when the interviewees were presented with the graph of a sampling distribution, the interviewees did not know what the graph represented nor did they understand its role in statistical hypothesis testing. They only thing familiar about this graph was the distribution. It was a normal distribution. However, the interviewees didn't really understand why it would be normally distributed. This reaction to the graph gives further information about the nature of the relationship between students' understandings of the procedures and the logic and concepts. Though the concept of a sampling distribution is central in statistical hypothesis testing, the students did not have an understanding that this was the case. This "gap" in their understanding of statistical hypothesis testing was filled by an understanding of the procedures.

It is interesting to note that the places where the interviewees relied on procedures to explain the logic and reasoning of statistical hypothesis testing were those where probability and inference were involved. Sampling distributions, rejection regions, levels of significance, and statistical tables are all concepts used to apply probability to determine the degree to which the sample is unusual under the null condition. In relying on probability, then, statistical hypothesis tests are useful as *inferential* methods. As with all inferential methods, statistical hypothesis testing deals with the variability and uncertainty associated with samples in order to draw conclusions about the population. Though there is evidence that, on some level, though the interviewees understood that there was uncertainty involved in statistical hypothesis testing, the interviewees tended to rely on rules and procedures to deal with that uncertainty. And, they did so without really appreciating the role of probability in those procedures. Thus, the relationship between student understandings of the procedures and of the logic and concepts is not very strong.

We now turn to an analysis of the relationship between introductory statistics students' understandings of the procedures and the uses of statistical hypothesis testing. For that, we will analyze student responses to the ***method and context*** items.

Understanding of the Procedures and of the Uses

The relationship between student understanding of the procedures and of the uses of statistical hypothesis testing is not as clear as that of the relationship of between student understanding procedures and the logic and concepts. In the explanations offered by the interviewees for their reasoning for the ***method and context*** items, there is not a great deal of evidence supporting the notion that the interviewees relied on procedures to answer the questions. This does not mean, however, that the interviewees used correct reasoning to choose correct answer choices. Their reasoning on these ***method and context*** items was limited by a lack of understanding of the "messiness" of real world data and of the way that statistical hypothesis tests deal with that "messiness" in drawing an inference about a population based on information from a sample. Because, as indicated above, the interviewees had knowledge of the procedures but exhibited a lack of understanding about the use of those procedures in "real world" settings, we know something about the relationship between the interviewees' understandings of the procedures and of the uses of statistical hypothesis testing.

In order to answer the ***method and context*** items, the introductory statistics students had to consider the context presented in each item. Given a set of potential research questions, they had to choose the one for which statistical hypothesis testing was an appropriate method of investigation. Given the mean and standard deviation of two samples of help line calls (with no other analyses), the students had to choose the most reasonable explanation for the differences. Given that the results of a statistical hypothesis test comparing two textbooks were statistically significant, the students had to determine which factors to consider in deciding whether or not to adopt the books. In order to answer each question correctly, the introductory statistics students had to understand that (1) "real world" data is messy; (2) that samples vary; and (3) that statistical hypothesis testing, while not a proof, is useful in helping to draw inference amidst that variability. In the analysis of the interviewees' explanations for their answer choices, it was found that they did not have strong, connected understandings of these concepts and ideas.

As indicated by the previous discussion, the interviewees did have knowledge of the procedures associated with statistical hypothesis testing. However, they did not have strong understandings of the logic and concepts, especially as related to the role of inference and probability. These concepts are central to an understanding of the role of context in the use of the method. Therefore, the explanations of the interviewees in the follow-up interviews provide further evidence that knowledge of the procedures of statistical hypothesis testing does not necessarily translate into an appreciation of the complexity of the context in which statistical hypothesis testing is employed. That is, the relationship between students' understandings of the procedures and of the uses of statistical hypothesis testing is not strong. An understanding of the uses of statistical hypothesis testing requires understanding of more than simply the procedures. It requires an appreciation of the complexity of the context and the role of statistical hypothesis testing in dealing with that complexity.

Summary

The analysis of the qualitative data with respect to research sub-question number 1 confirms the findings of the quantitative phase: the relationship between introductory statistics students' understandings of the procedures and of the logic, concepts, and uses of statistical hypothesis testing is not strong. The follow-up interviews, however, extended this finding and provided more information on why that relationship is not strong. The data supports the claim that just because a student knows the procedures associated with statistical hypothesis testing, he/she does not necessarily have an understanding of or appreciation for the role of *probability* and *inference* in statistical hypothesis testing. Even with an understanding of the procedures, introductory statistics students do not fully appreciate the value of statistical hypothesis testing as an inferential method which relies on indirect reasoning and probability to draw an inference about a population based on information from a sample.

Summary and Conclusions: Qualitative Phase

Overall, the results and analyses of the qualitative data confirm that which was found in the quantitative phase: Introductory statistics students do not have strong, connected understandings of the method and its uses and they are not very articulate in explaining their understandings of these ideas. The data collected in the qualitative phase provides more insight into why this is the case. In particular, the qualitative data (1) highlights the fact that introductory statistics students do not understand the role of indirect reasoning in statistical hypothesis and (2) provides evidence that introductory statistics students do not have strong understandings of and are uncomfortable with uncertainty, variability, probability, and inference. This struggle permeates through virtually every aspect of introductory statistics students' reasoning about the various components of statistical hypothesis testing.

From the *Adding It Up* (Kilpatrick, Swafford, & Findell, 2001) stranded perspective of proficiency, it can be said that introductory statistics students have fairly strong degrees of *procedural fluency* with respect to statistical hypothesis testing. They can perform the steps and are able to solve well-defined, traditional statistical hypothesis testing problems. However, there are gaps in the degree to which introductory statistics students have *conceptual understanding*, *adaptive reasoning*, *strategic competence*, and *productive disposition* in relation to statistical hypothesis testing. And, as indicated by the qualitative data, in particular, these competencies are limited by a lack of understanding of uncertainty, variability, probability, and inference.

Introductory statistics students do not understand that statistical hypothesis tests use sampling distributions and probability to determine whether a sample is unusual, conditioned on the null hypothesis. In fact, they do not understand what sampling distributions are. This lack of understanding speaks to the degree to which introductory statistics students have *conceptual understanding* and *adaptive reasoning* with respect to statistical hypothesis testing. They do not have deep, connected understandings of the overall logic and reasoning nor do they understand why concepts such as sampling distributions are fundamental to the overall reasoning. They do not understand the use of probability to determine whether the sample is unusual and to ultimately draw an inference about the population.

This lack of understanding of the overall logic of statistical hypothesis testing is complicated by the lack of appreciation introductory statistics students have of the variance that exists within and among samples from the same population. The data collected in the qualitative phase indicated that introductory students' reasoning is limited by a superficial understanding of the uncertainty and variability associated with sampling. This weak understanding impacts the ways in which introductory students interpret sample data and the results from a statistical hypothesis test as well as the ways in which they use that information to make decisions in "real world" contexts. Because introductory statistics students do not have strong understandings of these concepts, they tend to draw inferences from samples and hypothesis tests that are invalid. In addition, they do not have a sense of which factors they should take into consideration when using the results to inform real world practice. These are all indications of a weak degree of *adaptive reasoning* with respect to statistical hypothesis testing as they do not know how the ideas and concepts connect so that they may justify the decisions they make.

Introductory statistics students are, however, able to determine, for the most part, when statistical hypothesis testing is useful to answer a research question. Given a situation, introductory statistics students are able to set up the hypotheses that would be used in a statistical hypothesis test. These skills indicate a degree of *strategic competence*. However, as indicated by the qualitative data, this competence is limited by the fact that introductory statistics students do not understand the value of statistical hypothesis testing as an *inferential* method. Because introductory statistics students do not fully appreciate the way in which statistical hypothesis tests deal with the variability that is present within and among samples, they do not understand that statistical hypothesis tests are most useful to answer research questions about large populations. This lack of appreciation speaks to the degree to which introductory statistics students have a *productive disposition* toward the method and it uses.

In summary, the results of the qualitative phase highlight components of statistical hypothesis testing for which introductory statistics students do not have strong understandings. Because introductory students do not have strong understandings of uncertainty, variance, probability, and inference, they are not able to see the part these concepts play in the method of statistical hypothesis testing. Introductory statistics students do not understand that indirect reasoning is employed by statistical hypothesis tests to draw inferences about populations using information for a sample. Without these understandings, introductory statistics students do not have an appreciation for the method and its uses. Hence, these are understandings that should be developed throughout the introductory statistics course and that should be the focus of instruction.

CHAPTER 6

DISCUSSION

The study presented in this dissertation used a mixed methods approach to explore introductory statistics students' understandings of statistical hypothesis testing. In the quantitative phase, a multiple-choice assessment and course exam were used to provide large scale information about student understanding of the method and its uses. In the qualitative phase, follow-up interviews were used to probe students' thinking about the items on the multiple-choice assessment. Overall, the data collected in this study provide information about the degree to which introductory statistics students understand the "big picture" of statistical hypothesis testing.

In this chapter, a discussion of the study will be presented. This discussion is broken into the following four sections: (1) summary of the results and overall conclusion of the study, (2) presentation of the contributions and implications of the study, (3) statement of the limitations of the study, and (4) suggestions for future research.

Summary of Results and Overall Conclusion

The study presented in this dissertation was designed to explore introductory statistics students' understandings of statistical hypothesis testing. In particular, this study addressed the following research question: What are the understandings of statistical hypothesis testing held by students who have completed an introductory course in statistics at a large university? More specifically, the following three research sub-questions were addressed:

1. *What is the relationship between introductory students' understanding of the procedures and the concepts, logic, and uses of statistical hypothesis testing?*
2. *What are the understandings that introductory students have of the overall logic and reasoning of statistical hypothesis testing?*
3. *What are introductory students' understandings of the relationship between the method of statistical hypothesis testing and the context in which it is employed?*

In order to address these research sub-questions, large scale quantitative and small scale qualitative data were collected and analyzed. The results of this analysis indicate that introductory statistics students do not have strong, connected understandings of the overall "big picture" of statistical hypothesis testing.

Research Sub-question Number 1

With regard to the first research sub-question, the qualitative and quantitative data indicated that the relationship between introductory statistics students' understandings of the procedures and the concepts, logic and uses of statistical hypothesis testing is not strong. In the quantitative phase, participants scored well on the course assessment (largely a measure of *procedural fluency*) but did not score well on the multiple-choice assessment (a measure of *conceptual understanding*, *strategic competence*, *adaptive reasoning*, and, to some degree, *procedural fluency*). In addition, the scores on the two assessments were not strongly correlated. These results indicate that, though introductory statistics students may be able to perform the steps required to solve well-defined, traditionally worded statistical hypothesis testing problems, it does not necessarily mean they have strong, connected understandings of the "big picture".

The data collected in the qualitative phase provided more insight into the *nature* of the relationship between introductory statistics students' understandings of the procedures and the concepts, logic, and uses of statistical hypothesis testing. As they explained their reasoning on the multiple-choice items, the interviewees tended to rely on statements of procedures to justify their answers.

A reliance on procedures was particularly evident in the interviewees' discussion of the ***logic and reasoning*** items. The interviewees referred to the rule "the null hypothesis is equal" to establish the null hypothesis but couldn't explain why the null hypothesis should be a statement of equality. The interviewees also found it difficult to explain the logic and concepts that support the transition from sample analysis to statement of a conclusion. They could explain procedures that involved formulas and statistical tables to determine whether or not the null hypothesis should be rejected, but they didn't know what those tables represent or why the formulas and tables are used. The interviewees did not understand how probability and/or sampling distributions are used in the process.

In addition, a reliance on procedural knowledge hindered the interviewees' ability to consider the role of context when using and interpreting the results of a statistical hypothesis test. This struggle was evident in the discussions surrounding the ***method and context*** items. The interviewees did not seem to understand or be comfortable with concepts such as variability, uncertainty, and inference. When these concepts arose in discussion, the interviewees struggled to talk about them. Instead they recited a set of rules.

Together the data collected from the quantitative and qualitative phases indicate that the relationship between introductory statistics students' understandings of the procedures and the concepts, logic, and uses of statistical hypothesis testing is not strong. In particular, students tend to rely on procedures when asked to reason about those parts of statistical hypothesis testing that involve inference, probability, and uncertainty.

Research Sub-question Number 2

With regard to the second research sub-question, the qualitative and quantitative data indicated that introductory statistics students do not have strong understandings of the logic and reasoning of statistical hypothesis testing. They do not have strong understandings of the role of indirect reasoning in the method nor do they understand the concepts that support that logic and reasoning.

In the quantitative phase, scores on the ***logic and reasoning*** items from the multiple-choice assessment were low. Particularly low performances were recorded for items classified in the **Analysis of the Sample**, **Decision Rule**, and **Conclusion** categories. Analysis of the frequencies with which various distractors were chosen indicated that the students did not have strong conceptual understandings of sampling distributions, statistical significance, and p-values. In addition, given information about a sample and/or the conclusion made in a statistical hypothesis test, the introductory statistics students did not draw valid inferences. Their answer choices indicated that they understood statistical hypothesis tests to be methods of proof and/or methods by which the "truth" of a hypothesis can be measured. Furthermore, their answer choices indicated that they believed that sample statistics provide direct measures of population parameters.

These results indicate that introductory statistics students do not have strong understandings of sample variability.

The data collected in the qualitative phase confirmed most of these findings and provided more insight into how introductory statistics students understand the logic and reasoning of statistical hypothesis testing. Overall, the explanations offered by interviewees did not evidence an understanding of the roles of indirect reasoning and inference in the method. They did not understand that statistical hypothesis tests are used to assess the degree to which the collected sample is *unusual* under an assumed null condition and that, as a result of this analysis, an inference about the population can be made.

In addition, the interviewees did not understand statistical significance nor were they comfortable with the concepts of probability, uncertainty, and inference. In fact, some of the interviewees did not even know how to interpret the words “inference”, “uncertainty”, and “statistically significant”. In their explanations of their thinking for items in which those words appeared, many of the interviewees struggled to talk about those words. Some of the interviewees even claimed that “(un)certainty” and “probability” are not concepts associated with statistical hypothesis testing. Furthermore, the interviewees did not have an understanding of the degree to which samples vary nor did they have an understanding of sampling distributions and their role in statistical hypothesis testing. These ideas are foundational to the logic and reasoning of statistical hypothesis testing and provide a means by which inferences can be made. Because they did not understand the logic and reasoning that supports the method, the interviewees did not appreciate the value of statistical hypothesis as an inferential method and often drew conclusions about the population directly from sample information. These results confirm, but extend, findings in the quantitative phase.

Contrary to the results reported in the quantitative phase, however, the results reported in the qualitative phase indicated that introductory statistics students do not necessarily think that statistical hypothesis tests prove a given hypothesis. On the multiple-choice assessment interviewee's answers indicated belief that statistical hypothesis tests are proofs. But the interview data suggested this was not the case. The interviewees understood that the decision to reject or fail to reject the null hypothesis is based on information from a sample and that, unless all members of the population are tested, the conclusion is not necessarily true. However, since the interviewees did not fully understand the logic and reasoning that supports the conclusion, they often incorrectly interpreted the results of a statistical hypothesis test. They believed that the level of significance indicates the degree to which the researcher may be certain that the "accepted" hypothesis is true.

Together, the information collected from the quantitative and qualitative phases indicates that introductory statistics students do not have strong understandings of the logic and reasoning of statistical hypothesis testing. In particular, these students do not understand the role of indirect reasoning and probability in the method nor do they understand the role of inference in the transition from sample analysis to the statement of the conclusion. Hence, they do not value statistical hypothesis testing as an *inferential* method useful for studying large populations. Ultimately, this lack of understanding impacts the way in which introductory statistics students interpret the conclusions of statistical hypothesis tests.

Research Sub-Question Number 3

With regard to the third research sub-question, the qualitative and quantitative data indicated that introductory statistics students do not have very strong understandings of the relationship between the method and the context of statistical hypothesis testing. In particular, they do not appreciate the “messiness” of real world data and do not have strong understandings of statistical hypothesis tests as inferential methods used to “make sense” of that “messy” data.

In the quantitative phase, scores on the ***method and context*** items were stronger than those for the ***logic and reasoning*** items. However, these scores were still low. In particular, low scores were reported for items classified in the **Implication for Practice** category. Answers to items in that category indicated that introductory statistics students did not necessarily have a sense of the factors that should be considered when using the results of a statistical hypothesis test to inform decisions made in the “real world”. In addition, answer choices for the ***method and context*** items indicated that introductory statistics students were reluctant to attribute variability in sample statistics to chance. These results indicate a lack of appreciation for the variability that exists in data as well as a lack of appreciation for the value of statistical hypothesis testing. In using a statistical hypothesis test to analyze data, a researcher is able to make inferences about the degree to which the results may have differed by chance. Further evidence of an appreciation for statistical hypothesis testing as an inferential method was evidenced in the **Recognizing Applicability** category. Though students understood that statistical hypothesis tests could be used to compare two populations, their answer choices indicated a preference for applying statistical hypothesis testing in situations that did not require an inference to be made.

The results of the qualitative phase confirmed those of the quantitative phase but provided insight into the struggles that introductory statistics students face in linking the statistical hypothesis testing method to the context in which it is employed. As was found in the analysis of interview data for the ***logic and reasoning*** items, interviewees were uncomfortable with the concepts of uncertainty and variability. Hence, they did not appreciate the value of statistical hypothesis testing as a means of making inferences about populations amidst the variability and uncertainty associated with samples. The interviewees thought that statistical hypothesis tests are best used when the population being studied is not too “broad”, so to speak. The interviewees were uncomfortable applying statistical hypothesis testing to research questions that addressed large populations when, in fact, these are the very populations statistical hypothesis testing is designed to address. In addition, as indicated in the quantitative phase, the interviewees were reluctant to attribute differences in samples to chance. Instead, the interviewees gave causal reasons for the differences and/or drew upon their own experiences to explain those differences. However, as indicated in the interview data for the ***logic and reasoning*** items, the interviewees did recognize that the results of a statistical hypothesis test do not necessarily prove a hypothesis and, thus, understood that action should not necessarily be taken based on the results of one study alone. They felt that before action is taken, further analysis should be done. However, they did not know what kind of analysis is appropriate in those circumstances.

Together, the results from the quantitative and qualitative phases indicate that introductory statistics students do not have strong understandings of the relationship between the statistical hypothesis testing method and the context in which it is employed. In particular, introductory statistics students do not have strong understandings of the variability associated with samples and they do not value statistical hypothesis testing as an inferential method. This lack of understanding impacts their reasoning about situations in which the method should be applied and their reasoning about the ways in which the results of a statistical hypothesis test should be used to inform "real world" decision making.

Conclusion

Taken together, the quantitative and qualitative phases of this study provide information about introductory statistics students' understandings of the "big picture" of statistical hypothesis testing. As is evident from the results presented above (and in Chapters 4 and 5) the two phases of the study informed each other. The quantitative phase was conducted on a large scale, and provided general information about the overall understandings of many introductory statistics students. The qualitative data confirmed and extended the findings of the quantitative phase. Through the data collected in the follow-up interviews, it was possible to gain more insight into student thinking than that which was provided in the quantitative phase. In addition, through the follow-up interviews, we saw that introductory statistics students are not very articulate in expressing their understandings of the concepts and ideas associated with statistical hypothesis testing.

Overall, the data collected indicates that introductory statistics students, who have completed a traditional course at a large university, do not have strong, connected understandings of the "big picture" of statistical hypothesis testing. Though these students can solve well-defined, traditionally worded, statistical hypothesis testing problems, they do not have strong understandings of the logic and reasoning of statistical hypothesis testing. In addition, they do not have strong understandings of the relationship between the method and the context in which it is employed. In particular, introductory statistics students do not understand the role of indirect reasoning in the method. They are uncomfortable with concepts such as probability, uncertainty, and variability and do not understand sampling distributions and their role in statistical hypothesis testing. Introductory statistics students do not value statistical hypothesis testing as an inferential method used to draw inferences from "messy" real-world data.

However, the results indicated that introductory statistics students *do* understand that statistical hypothesis tests are not proofs and that the conclusion reached is not necessarily correct. In addition, introductory statistics students understand statistical hypothesis tests to be useful in comparing two groups.

From the perspective of proficiency offered by *Adding It Up* (Kilpatrick, et al., 2001), these results indicate that, at the end of a traditional, introductory statistics course at a large university, students have fairly high degrees of *procedural fluency* with respect to statistical hypothesis testing. However, they do not have high degrees of *conceptual understanding*, *strategic competence*, *adaptive reasoning*, or *procedural fluency* (at least with respect to understanding the *value* of the method). In addition the relationship between and among these "strands" is weak.

Introductory statistics students, however, are not without any understanding. They do have some understandings that are useful and that can be built upon. It may be the case, that over time, with more exposure to statistical hypothesis testing and to statistics in general, these understandings will develop. However, introductory statistics students do not have these understandings when they complete a first, introductory course in statistics.

Contributions and Implications

The results of this study advance a growing body of knowledge about student understanding of statistical hypothesis testing. Additionally, the instrumentation developed in this study provides a guide for assessment of that understanding. Taken together, the results and products of this study have implications for instructional design and student assessment in introductory statistics courses.

Contributions

A major contribution of this study is the fulfillment of a gap in the literature on student understanding of statistical hypothesis testing. Few studies examine understandings of the overall "big picture" of statistical hypothesis testing. The literature includes some studies that examine student understanding of the *procedures* (Aquilonius, 2005; Evangelista & Hemenway, 2002; Hong & O'Neil, 1992; Link, 2000) and of the *components* (Hong & O'Neil, 1992; Krauss & Wassner, 2002; Lane-Getaz, 2007; Lipson, Kokonis, & Francis, 2003; Mittag & Thompson, 2000; Wilkerson & Olson, 1997) of statistical hypothesis testing. However, only two studies have been conducted that examine understandings of statistical hypothesis testing beyond procedural knowledge (Aquilonius, 2005; Liu, 2005) and these studies are limited in that they (1) did not examine understandings of all aspects of the "big picture" of statistical hypothesis and (2) were done with only small numbers of participants. Therefore, large scale study of student understanding of the overall "big picture" of statistical hypothesis testing was needed to advance the field. This study addresses that need. It provides large scale information on the degree to which students enrolled in a traditional, introductory statistics course at a large university understand the overall "big picture" of statistical hypothesis testing. Furthermore, in order to support the large scale analysis of student understanding, this study included analysis of data collected in follow-up interviews designed to further explore student thinking on a small scale.

The results of this study confirm many of the findings reported in the literature on student understanding of statistical hypothesis testing. Like the students in Aquilonius' (2005) study, the participants in the present study were able to implement the procedures associated with statistical hypothesis testing but did not have strong understandings of the concepts that supported those procedures. The students in both studies did not readily refer to sampling distributions or probability in their reasoning, did not understand *p*-values, and relied on well-rehearsed rules to justify their reasoning.

In addition, like the teachers in Liu's (2005) study, the participants in the present study did not understand the role of indirect reasoning in statistical hypothesis testing. Neither group of participants understood that statistical hypothesis tests are used to determine the degree to which the collected sample is *unusual* under the assumed null condition. And, neither group had strong understandings of the relationship between the method and the context. When presented with sample information, participants in both studies relied on their own personal experience to judge whether or not the null hypothesis should be rejected, rather than appealing to statistical hypothesis testing to make that decision. And, both groups were reluctant to attribute variability in sample information to chance. Instead, they attributed the difference to an actual difference in the population or to some other causal event.

Though the findings of the present study confirm those of Aquilonius' (2005) and Liu's (2005) studies, these findings also provide *additional* information about student understanding useful in advancement of the field. Because the present study was done on a large scale with introductory statistics students at a large university, we now have evidence that many of these students have incomplete understandings of the method and its uses. In addition, as a result of the present study, we now have more information about student understanding of other aspects of the overall "big picture" of statistical hypothesis testing. We know more about student understanding of the logic and reasoning of statistical hypothesis testing and we know more about their understanding of the relationship of the method to the context in which it is employed. Specifically, we now have evidence that introductory statistics students do not have strong understandings of the variability and uncertainty associated with samples, they do not understand sampling distributions and their role in statistical hypothesis testing, and they do not understand the way in which statistical hypothesis tests draw inferences about a population based on information from a sample. Furthermore, we now have evidence that introductory statistics students do not value statistical hypothesis testing as an inferential method that is reliant on probability theory and that is useful for making decisions in “real world” contexts.

The results of the present study also confirm, but extend, what is known about student understanding of the individual *components* of statistical hypothesis testing. In the present study, we saw evidence that introductory statistics students do not understand the *components* of statistical hypothesis testing. These findings are consistent with that which has been reported in the literature on student understanding of those components. As a result of this study, however, we now see how difficulty understanding the *components* can limit the degree to which introductory statistics students understand statistical hypothesis testing overall.

In addition to a contribution of results that advance the field, the present study makes a contribution to field of statistics education in the creation of products useful to the study of student understanding of statistical hypothesis testing. Prior to this study, the literature did not contain a framework that outlines understandings important to a strong, connected understanding of statistical hypothesis testing. Additionally, none of the studies in the literature had reported on a multiple-choice assessment of *those* understandings. In order to assess student understanding of the concepts, logic, and uses of statistical hypothesis testing on a large scale, it was necessary to create such a framework and instrument. Thus, *initial* versions of a Framework for Assessing Understanding of Statistical Hypothesis Testing and a multiple-choice assessment of student understandings of the concepts, logic, and uses of statistical hypothesis testing were created. Though the Framework and multiple-choice assessment are not perfected, they are products of this study that may be used and/or refined to be used in future studies of instruction and student understanding. In offering these products, the present study makes a contribution to the field of statistics education.

Implications

Taken together, the results of the studies on student understanding of statistical hypothesis testing (past and present) have implications for instruction on statistical hypothesis testing. The results indicate that instruction should focus more attention on the development of student understanding of the overall “big picture” of statistical hypothesis testing, in addition to mastery of the procedures. In particular, more emphasis should be placed on the development of strong understandings of the logic of indirect reasoning and its role in statistical hypothesis testing. In addition, more emphasis should be placed on the development of an understanding of statistics as a field of study different from mathematics. In so doing, care should be taken to help introductory statistics students become more comfortable with uncertainty and variability. They should come to appreciate statistical hypothesis testing as an *inferential* method used to “manage” variability in samples. Instruction should help students to understand sampling distributions as necessary components to the overall logic and reasoning of statistical hypothesis testing. Finally, instruction should provide students with plenty of opportunities to apply statistical hypothesis testing to ill-defined problems that involve “real world” data. In summary, instruction should provide students with activities and tasks that help them to build understandings of the “big picture” of statistical hypothesis testing such that they appreciate the value of the method and its uses.

In addition, the Framework and multiple-choice assessment developed in this study offer implications for instruction in statistical hypothesis testing. These products of the study highlight important understandings of the method and its uses that have been advocated for by leaders in the field (e.g. ASA, 2005; Cobb, 1992; Garfield & Chance, 2000; Moore, 1997; Snee, 1999). As such, they are useful in guiding instruction about and/or assessing student understanding of the concepts, logic, and uses of statistical hypothesis testing.

Limitations

Though the study described in this dissertation contributes to the field of statistics education, there are limitations to the conclusions that can be drawn from the findings. As is the case for all studies in the field of education, the conclusions of this study are limited by the overall design.

The participants in this study were enrolled in a traditional, introductory statistics course at a large university. Though there is evidence to suggest that statistical instruction in introductory statistics courses on the whole is very similar across universities (Garfield, Hogg, Schau, & Whittinghill, 2002; Shaughnessy, 1992), caution should be exercised in generalizing the results of this study to other populations.

Though comparison of performance on the third course exam of the participants in the study with the overall performance of students enrolled in STAT 100 indicated that the sample was representative of the overall population (see Chapter 4), it should be noted that the participants in this study were volunteers. Therefore, self-selection bias should be taken into consideration when attempting to generalize the results to other populations.

In addition, because this study was not an investigation of instruction, classroom observations were not conducted. We do not know how each instructor presented the material and/or talked about the concepts and ideas involved. Furthermore, we do not know how individual instructors graded each of the third course exams. In essence, we do not know the effect of the instructor and/or course design on student performance. The approach used by other instructors at other universities may be different and the extent to which concepts are addressed may be different from university to university and from course to course. We do know, however, that instruction in this course was lecture-based and guided by a traditional textbook. In addition, we know that assessments in this course were composed of well-defined, traditional statistical hypothesis testing problems. Hence, to the extent that that other university courses are taught in similar lecture-style formats with a focus on procedures, the results of this study *inform* our understanding of student learning about statistical hypothesis testing from introductory statistics courses at large universities.

This study was also limited by its use of a multiple-choice assessment to gather large scale data on student understanding of statistical hypothesis testing. Though items from the instrument were piloted and modifications were made, the final version of the instrument was *not* piloted. This instrument has not been subjected to tests of reliability or validity. In fact, as indicated by student responses in the follow-up interview, students often did not interpret the language used in the instrument to mean what was intended when the instrument was developed. Additionally, because the instrument used a multiple-choice format, the participants were provided with potential answers. Had the items been presented in a “free response” format, the participants may have offered answers quite different from those provided on the assessment. Given four answer choices, however, the participants were able to use test taking strategies to eliminate distractors and make "educated" guesses as to which answer choice is correct, whether or not it was the answer they had in mind upon reading the question. These issues highlight the limitations associated with the use of multiple-choice assessments in general. Thus, the mixed methods approach used in this study was useful to gain more insight into student thinking and to provide triangulation of data sources.

Finally, there are limitations associated with the results reported in the qualitative phase. Though care was taken to choose interviewees who were representative of the various performance patterns on quantitative instruments, the interviewees chosen represent a small fraction of the overall population of introductory statistics students. Each of the interviewees is an individual with unique characteristics and unique understandings. Therefore, caution should be exercised in generalizing the results of this phase of the study to other populations. However, because the interviewees *were* chosen to maximize the potential for variability that could exist in student responses, their responses do provide *some* insight into the ways introductory statistics students may or may not understand statistical hypothesis testing.

Future Research

Given the results and limitations of this study, there are several follow-up studies that can be done to build on this work and advance the field. Given the results reported in this and other studies, it is of interest to study instruction itself. We have evidence that introductory statistics students do not have strong, connected understandings of statistical hypothesis testing and we know where students have difficulty in developing those understandings. These results imply that instruction should be enhanced to focus on those areas for which introductory statistics students are weak.

As a “next step” it is important to study the *current* state of instruction. Studies of instruction in introductory statistics courses at large universities would provide information on how statistical hypothesis testing is addressed in these courses. Because instruction at large universities is often guided by textbooks, a textbook analysis would provide useful information about how the concept is treated and dealt with in the context of traditional courses. In fact, a textbook analysis of a variety of textbooks (not just those typically used in university settings) would be useful. It is of interest to see how different textbooks discuss the logic, reasoning, and uses of statistical hypothesis and/or to what extent they focus on the procedures relative to other aspects of the “big picture”. Additionally, a discourse analysis of classroom discussions would provide information about how statistical hypothesis testing is presented and discussed as well as how the students respond. Such studies would provide valuable information about the relationship between instruction and student learning and would inform future design of instruction at large universities.

As instruction is modified, it is important then to study the impact of those modifications on student understanding. Research on instruction should study the learning process itself so that educators and researchers develop stronger understandings how various instructional activities promote (or limit) the development of strong understandings of the “big picture” of statistical hypothesis testing. Here again, discourse and/or textbook analysis would be useful to study the impact of those enhancements to instruction.

The Framework and multiple-choice assessment developed in the present study should be refined. With further refinement, these objects can be useful in evaluating the success of various instructional designs on the development of student understanding of the concepts, logic, and uses of statistical hypothesis testing.

Ultimately, the line of inquiry addressing student understanding of statistical hypothesis testing should be applied to studies of *teacher* understanding of the “big picture” of statistical hypothesis testing. As was the case in this study, many introductory statistics teachers are trained in mathematics. They, too, should develop strong, connected understandings of statistical hypothesis testing (and of statistics, in general). Research on the development of teacher understanding of these ideas therefore, is of value.

Ultimately, a line of inquiry that includes studies both of instructional design in introductory statistics course and of the development of teacher knowledge about statistical hypothesis testing have the potential to improve the degree to which students complete an introductory statistics course with strong, connected understandings of the statistical hypothesis testing and its uses. And, given today’s data driven world, these understandings are of value.

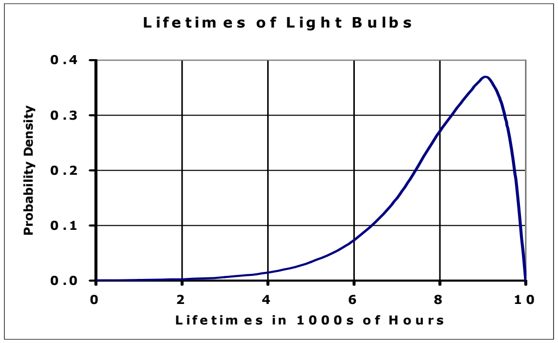
APPENDIX A

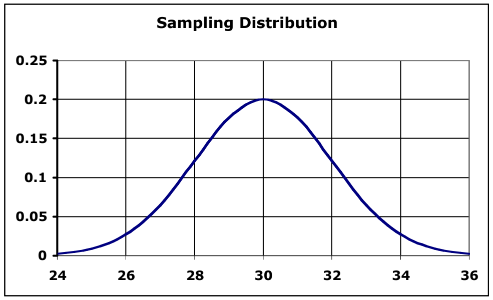
REASONS FOR DISTRACTORS

Item and Code	Code Answer	Distractor Assesses
Which of the following questions is ***most likely*** to be answered by a study that requires statistical hypothesis testing? a. Do athletes have a lower GPA than other students? b. What equation predicts a student's freshman GPA from his/her SAT score? c. What are typical costs for full-time resident students in U. S. colleges? d. Do the 12:00 noon sections of STAT 100 perform better than the 2:00 p.m. sections this semester?	RA (1.1, 1.2, 1.3) A	All distractors address indication from the research that individuals do not know when to use hypothesis testing (Liu, 2005) and/or which test to use (Aquilonius, 2005) d. Assesses whether students understand that statistical hypothesis testing is not necessary when it is possible to obtain information from the entire population. Assesses whether students value statistical hypothesis testing as an inferential method used to make inferences about a population based on information from a sample.
Which of the following actions is the ***most important first step*** in designing a statistical hypothesis test to answer the question: *Are out-of-state students more successful at the state university than students who are in-state residents?* a. Agree on a statistical test to compare the groups. b. Agree on a sample size from each population. c. Identify available statistical software. d. Agree on a way of measuring student success.	RA (1.1) D	All distractors address indication from research that individuals do not know when or how to use hypothesis testing (Liu, 2005; Aquilonius, 2005) Although all distractors present important decisions that must be made, a statistical hypothesis test cannot be under consideration unless that question can be quantified.
Which of the following statements is the ***best justification*** for using a statistical hypothesis test to answer the question: *Are female students as successful as male students in college mathematics?* a. It allows you to talk about uncertainty associated with your decision.	RA (2) A	b. Assesses whether students think that statistical hypothesis testing provides a proof for either the null or alternative hypotheses. This understanding might arise from a belief that one sample proves a statement for a population (Bady, 1979; Klaczynski, 2000; Kuhn & Dean, 2004; Wason, 1960), from reasoning according to the representativeness heuristic (Kahneman

b. It allows you to use only a sample of students to prove something for all students. c. It allows you to calculate means and use statistical tables to answer the question. d. It allows you to find and prove the answer using mathematical calculation.		and Tversky, 1972, 1982) and/or from difficulty in managing the tension between sample representativeness and sample variability (Bruce, 2006; Reading & Shaughnessy, 2004) c. Might be the kind of response from someone who thinks of the test only as a procedure consisting of calculations (Aquilonius, 2005). d. Might be the kind of response from someone who thinks of the test as a procedure (Aquilonius, 2005). Also assesses whether students think that statistical hypothesis testing provides a proof for either the null or alternative hypotheses (see justification for distractor b, above).
A researcher would like to establish that freshmen in the humanities have higher SAT scores than freshmen in the sciences. Which of the following ***null hypotheses*** should be tested? The mean SAT score of humanities students is … a. greater than that of science students. b. greater than or equal to that of science students. c. less than or equal to that of science students. d. less than that of science students.	GH (1) C	All distractors assess whether individuals confuse the inequality statements when establishing the null and alternative hypotheses (Link, 2002)
A local dry cleaner uses *Stain Away* stain remover, which is known to remove 85% of stains. A new, more expensive product *Erase Away* claims to be more effective than *Stain Away*. Which set of hypotheses should be used to test whether the dry cleaner should consider switching? a. H_0: *Erase Away* removes more than 85% of stains. H_1: *Erase Away* removes no more than 85% of stains. b. H_0: *Erase Away* removes 85% of stains.	GH (1, 2, 3) D	All distractors assess whether individuals confuse the inequality statements when establishing the null and alternative hypotheses (Link, 2002)

H_1: *Erase Away* does not remove 85% of stains. c. H_0: *Erase Away* removes more than 85% of stains. H_1: *Erase Away* removes less than 85% of stains. d. H_0: *Erase Away* removes no more than 85% of stains. H_1: *Erase Away* removes more than 85% of stains.		
To test the hypothesis that private schools are more effective than public schools, researchers plan to compare mean starting salaries of private and public school graduates. But they cannot agree on whether to test the results at a significance level of 0.10 ($\alpha = 0.10$) or at a significance level of 0.05 ($\alpha = 0.05$). What effect will using 0.10 rather than 0.05 have on the study? a. Using 0.10 will result in a greater chance that they will incorrectly retain the null hypothesis. b. Using 0.10 will result in a greater chance that the null hypothesis is actually true. c. Using 0.10 will result in a greater chance that they will incorrectly reject the null hypothesis. d. Using 0.10 will result in a greater chance that the alternative hypothesis is actually true.	DR (1) C	This item assesses student understanding of significance level. The distractors assess whether students understand significance level as the probability of making a Type I error, if the null is, in reality, correct. Individuals do not have a good understanding of *p*-value and interpret it to be 1. The probability that the null is/is not true 2. The probability that the alternative is/is not true (Lane-Getaz, 2007; Haller & Krauss, 2002; Nickerson, 2000) Interpretation of *p*-value relates to interpretation of significance level. Individuals believe 1. That alpha is the probability of making a Type I error 2. That alpha is the probability of making a Type I error if the study is repeated over time with the same alpha level 3. That alpha is the probability that if one has rejected the null, he/she has made a Type I error (Nickerson, 2000;) The distractors build off of these notions in assessing student understanding of significance level.
When *Consumer Reports* studied response times for a random sample of 60 computer help-line calls, they found a mean of 15 minutes and standard deviation of 4.5 minutes. After hearing complaints about decline in service, they repeated the study (again using a sample of 60 calls) and found a mean response time of 16.5 minutes and standard deviation of 6.0 minutes.	CS (1, 2) C	a. Assesses whether students are reasoning according to the representativeness heuristic (Kahneman and Tversky, 1972, 1982) and/or are struggling to manage the tension between sample representativeness and sample variability (Reading & Shaughnessy, 2004; Bruce, 2006). Additionally, he/she many be

What is ***the most plausible interpretation*** of the difference between the two study results? a. Because the second study showed a higher mean, that study must have only looked at computer help-lines that received a lot of consumer complaints. b. The increase in mean response time confirms a decline in services by computer help-lines. c. The observed difference in mean response times is quite possibly due to chance variation. d. The increase in standard deviation is the reason for the increase in mean response time.		reasoning according to personal beliefs (Klaczynski, 2000; Kuhn and Dean, 2004) while misunderstaning the notion of sample bias (Konold & Higgins, 2003; Watson & Moritz, 2000) b. Assesses whether students think that one sample proves a statement for a population (Bady, 1979; Klaczynski, 2000; Kuhn & Dean, 2004; Wason, 1960). d. Assesses students' understanding of variability and its' relationship to summary statistics.
The typical distribution of usable lifetimes for light bulbs is shown in the following sketch. It's clearly not a normal distribution. However, when doing a test to compare the mean lifetimes of two light bulb brands (using large samples of both brands), a statistician used a normal distribution to find the *p*-value of the difference.  Which of the following statements ***best explains why*** the test for significance involves normal probabilities, rather than probabilities from the light bulb lifetime	AS (2) A	b. Assesses whether students confuse the sampling distribution of the statistic with the distribution of values in a sample (delMas, Garfield, & Chance, 2004; Saldanha & Thompson, 2002) c. Assesses whether students have a procedural as opposed to a conceptual understanding of sampling distributions and their use in statistical hypothesis testing. Indicates a superficial understanding. d. Assesses whether students confuse the sampling distribution of the statistic with the distribution of values in a sample (delMas, Garfield, & Chance, 2004; Saldanha & Thompson 2002)

distribution? a. The distribution of the difference of means of large samples is always approximately normal. b. The distribution of values in large samples is always approximately normal. c. Values of the standard normal probability distribution are always given in reference tables. d. The distribution of differences of values in two samples is always approximately normal.		
Tests show that fuel efficiency of cars in the current model year averages 30 miles per gallon. A test of 100 new car models gave mean fuel efficiency of 31.5 miles per gallon. To see whether it is correct to claim that the new car models are more fuel-efficient than those in the current model year, a researcher constructed the sampling distribution of average fuel efficiencies of many samples of 100 current model cars in the following graph.  Which of these conclusions is ***best supported*** by the graph above? a. The difference in fuel efficiency of current and new model cars is not statistically significant. b. Half of current and new model cars have fuel-efficiencies below 30 miles per gallon.	AS (2, 3) A	b and d. Assess whether students confuse the sampling distribution of the statistic with the distribution of values in a sample (delMas, Garfield, & Chance, 2004; Saldanha & Thompson, 2002)

c. The difference in fuel efficiency of current and new model cars is statistically significant. d. Nothing can be concluded. The graph should be centered at 31.5.		
To test the effectiveness of a new method of teaching reading, researchers used the new method with a class of 35 second-grade students and found that 70% of those students were then reading above grade level. In a typical year, 50% of second-grade students are reading above grade level. In order to test the significance of the new program effect, researchers calculated the test statistic $z = \frac{0.7 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{35}}} \approx 2.37$ What is the ***best explanation*** of what the researchers learn by using a statistical table to find a *p*-value for the test statistic 2.37? a. The *p*-value tells the probability that the new teaching method results in a 20% gain in the number of students reading above grade level. b. The *p*-value tells the probability that the new teaching method does not result in a 20% gain in the number of students reading above grade level. c. The *p*-value tells the probability of getting the observed results, if the new program does result in better reading skill. d. The *p*-value tells the probability of getting the observed results, if the new program does not result in better reading skill.	AS (3) D	This item assess whether students understand *p*-value as the probability of the observed sample conditioned on the null hypothesis. Individuals do not have a good understanding of *p*-value and interpret it to be 1. The probability that the null is/is not true 2. The probability that the alternative is/is not true (Lane-Getaz, 2007; Krauss & Wassner, 2002; Nickerson, 2000) The distractors assess whether students have these, or related, misunderstandings of *p*-value.
In 1950 the mean IQ of undergraduates at a university was 110. To test the hypothesis that students today are smarter, a study of 500 current students found a mean IQ of 120. The difference between the two means is significant at the 0.05 level. ($\alpha = 0.05$) Which of the following statements is necessarily true?	C (1, 2, 3) D	a. Assesses whether students think that statistical hypothesis testing provides a proof for either the null or alternative hypotheses. This understanding might arise from a belief that one sample proves a statement for a population (Bady, 1979; Klaczynski, 2000; Kuhn & Dean, 2004; Wason, 1960), from reasoning according to the representativeness heuristic (Kahneman

<table>
<tr>
<td>a. Undergraduates at the university today are smarter than those in 1950.
b. The claim that undergraduates today are not smarter than those in 1950 is true with a probability less than 0.05.
c. The claim that undergraduates today are smarter than those in 1950 has been established with 95% certainty.
d. If undergraduates today are no smarter than those in 1950, the probability of the observed mean IQ is less than 0.05.</td>
<td></td>
<td>and Tversky, 1972, 1982) and/or from difficulty in managing the tension between sample representativeness and sample variability (Bruce, 2006; Reading & Shaughnessy, 2004)

b, c. Assesses whether students are interpreting statistical significance as the probability of either the null or alternative hypothesis being true. Relates to misunderstandings of *p*-value (Lane-Getaz, 2007; Krauss & Wassner, 2002; Nickerson, 2000)</td>
</tr>
<tr>
<td>A study tested the claim that: *Transfer students are less successful at the state university than students admitted as first time freshmen*. Results showed a difference in first semester grade point averages that is significant at the 0.05 level. Information from samples of transfer and first time freshmen is shown in the table below.

<table>
<tr><th></th><th>Transfer Admits</th><th>Freshman Admits</th></tr>
<tr><td>**n**</td><td>50</td><td>50</td></tr>
<tr><td>**mean gpa**</td><td>2.5</td><td>2.8</td></tr>
</table>

What is the ***most reasonable inference*** about the population of all first semester students that can be drawn from this information?

a. There are equal numbers of transfer and first time freshman students on campus.
b. The mean first semester GPA of all freshman admits is 0.3 greater than that of all transfer admits.
c. It is unlikely that the first semester GPA of all transfer admits equals that of all</td>
<td>C
(1, 2)

C</td>
<td>a, b, d. Assess whether students think that one sample proves a statement for a population (Bady, 1979; Klaczynski, 2000; Kuhn & Dean, 2004; Wason, 1960;) and/or assesses whether individuals are reasoning according to the representativeness heuristic (Kahneman and Tversky, 1972, 1982) and are struggling to manage the tension between sample representativeness and sample variability (Bruce, 2006; Reading & Shaughnessy, 2004)</td>
</tr>
</table>

freshman admits. d. The mean first semester GPA of all University students is $\frac{2.5+2.8}{2}$ or about 2.65.		
In an educational study the mean test score of students studying from a new, experimental textbook A was greater than that for students studying from the previously used, traditional textbook B, with significance at the α = 0.05 level. What action in response to that result ***makes most sense*** to you? a. Compare the mean scores to see if the difference is great enough to merit the cost of new books. b. Schools should adopt textbook A because its use leads to significantly better learning. c. Re-analyze the data to see if the difference in means is significant at the 0.10 level. d. Take no action until the study is repeated, because the difference in scores could be due to chance.	IP (1, 2, 3) A	This item assesses whether students understand that statistical significance does not always imply practical significance and that students recognize the need to examine effect size, at least informally. b. Assesses whether students think that statistical hypothesis testing provides a proof for either the null or alternative hypotheses. This understanding might arise from a belief that one sample proves a statement for a population (Bady, 1979; Klaczynski, 2000; Kuhn & Dean, 2004; Wason, 1960), from reasoning according to the representativeness heuristic (Kahneman and Tversky, 1972, 1982) and/or from difficulty in managing the tension between sample representativeness and sample variability (Bruce, 2006; Reading & Shaughnessy, 2004) c. Assesses whether students understand significance level. (See item number 11) d. Assesses whether students understand that role of statistical hypothesis testing in establishing a means for decision making based on evidence collected from *one*, usually expensive study. This distractor was written to reflect the findings that 1. teachers want to take another sample to "prove" the statement (Liu, 2005) 2. student don't think that one sample gives enough information (Jacobs, 1999) 3. individuals generally misunderstand sampling distributions (delMas, Garfield, & Chance, 2004; Saldanha & Thompson, 2002) and their role in hypothesis testing (Liu, 2005)

To evaluate a new computer-based approach to teaching pre-calculus, 200 volunteers among the 1000 pre-calculus students took the course on-line. At the end of the semester the mean final exam scores were 83.5 for the on-line students and 83.1 for the other students, a difference that proved to be significant at the 0.05 level. If you were unhappy with the resulting recommendation that the course be taught on-line to all students, which of the following critiques is ***least likely*** to be effective in challenging the recommendation? a. While the difference in means may be statistically significant, it is very small in practical terms. b. The study did not use random assignment of subjects to treatment groups. c. The experimental group was too small—only 200 out of 1000 students. d. A previous experiment on the calculus course did not show positive results for on-line instruction.	IP (1, 2) D	This item assesses whether students understand (1) common areas for which to critique a study and (2) that statistical significance does not always mean practical significance. Distractors a, b, and c are all legitimate challenges.

APPENDIX B

MULTIPLE-CHOICE ASSESSMENT

STAT 100 Multiple Choice Survey Cover Sheet

This survey contains questions about hypothesis testing. Please read each question and answer choice carefully before choosing the best option. Your answers to the questions will not affect your grade but will be used in a large-scale study of student understanding of statistical hypothesis testing, which may ultimately inform instruction for future students.

If you have not completed the appropriate consent form, please ask your instructor for the form so that you may participate in the study. You may choose to participate in the survey phase only or you may choose to participate in both the survey and follow-up interview phases. In either case, you must have completed the consent form.

Thanks so much for your participation! ☺

Name (print):

____

Name (sign):

____

Email:

____

Multiple Choice Survey

After reading the question and each answer choice, circle the best response to each of the following. You may write on this survey.

Researchers conduct many studies that describe and compare traits of university students. Items 1 - 6 ask about design, analysis, and interpretation of statistical studies to answer common questions.

1. Which of the following questions is ***most likely*** to be answered by a study that requires statistical hypothesis testing?

a. Do athletes have a lower GPA than other students?

b. What equation predicts a student's freshman GPA from his/her SAT score?

c. What are typical costs for full-time resident students in U. S. colleges?

d. Do the 12:00 noon sections of STAT 100 perform better than the 2:00 p.m. sections this semester?

2. Which of the following actions is the most ***important first step*** in designing a statistical hypothesis test to answer the question: *Are out-of-state students more successful at the state university than students who are in-state residents*?

a. Agree on a statistical test to compare the groups.

b. Agree on a sample size from each population.

c. Identify available statistical software.

d. Agree on a way of measuring student success.

3. Which of the following statements is the ***best justification*** for using a statistical hypothesis test to answer the question: *Are female students as successful as male students in college mathematics*?

a. It allows you to talk about uncertainty associated with your decision.

b. It allows you to use only a sample of students to prove something for all students.

c. It allows you to calculate means and use statistical tables to answer the question.

d. It allows you to find and prove the answer using mathematical calculation.

4. A researcher would like to establish that freshmen in the humanities have higher SAT scores than freshmen in the sciences. Which of the following ***null hypotheses*** should be tested?

The mean SAT score of humanities students is …

a. greater than that of science students.

b. greater than or equal to that of science students.

c. less than or equal to that of science students.

d. less than that of science students.

5. In 1950 the mean IQ of undergraduates at a university was 110. To test the hypothesis that students today are smarter, a study of 500 current students found a mean IQ of 120. The difference between the two means is significant at the 0.05 level. ($\alpha = 0.05$)

Which of the following statements is necessarily true?

a. Undergraduates at the university today are smarter than those in 1950.

b. The claim that undergraduates today are not smarter than those in 1950 is true with a probability less than 0.05.

c. The claim that undergraduates today are smarter than those in 1950 has been established with 95% certainty.

d. If undergraduates today are no smarter than those in 1950, the probability of the observed mean IQ is less than 0.05.

6. A study tested the claim that: *Transfer students are less successful at the state university than students admitted as first time freshmen.* Results showed a difference in first semester grade point averages that is significant at the 0.05 level. Information from samples of transfer and first time freshmen is shown in the table below.

	Transfer Admits	**Freshman Admits**
n	50	50
mean GPA	2.5	2.8

What is the ***most reasonable inference*** about the population of all first semester students that can be drawn from this information?

a. There are equal numbers of transfer and first time freshman students on campus.

b. The mean first semester GPA of all freshman admits is 0.3 greater than that of all transfer admits.

c. It is unlikely that the first semester GPA of all transfer admits equals that of all freshman admits.

d. The mean first semester GPA of all University students is $\frac{2.5+2.8}{2}$ or about 2.65.

~~~~~~
~~~~~~

Before deciding to make costly purchases, consumers often read reports of product quality. Items 7 - 10 ask about design, analysis, and interpretation of statistical product quality studies.

7. A local dry cleaner uses *Stain Away* stain remover, which is known to remove 85% of stains. A new, more expensive product *Erase Away* claims to be more effective than *Stain Away*. Which set of hypotheses should be used to test whether the dry cleaner should consider switching?

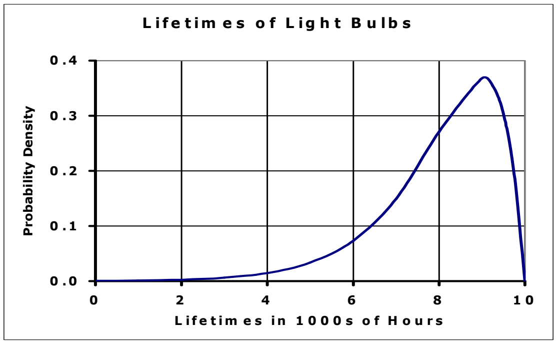
a. H_0: *Erase Away* removes more than 85% of stains.
   H_1: *Erase Away* removes no more than 85% of stains.

b. H_0: *Erase Away* removes 85% of stains.
   H_1: *Erase Away* does not remove 85% of stains.

c. H_0: *Erase Away* removes more than 85% of stains.
   H_1: *Erase Away* removes less than 85% of stains.

d. H_0: *Erase Away* removes no more than 85% of stains.
   H_1: *Erase Away* removes more than 85% of stains.

8. The typical distribution of usable lifetimes for light bulbs is shown in the following sketch. It's clearly not a normal distribution. However, when doing a test to compare the mean lifetimes of two light bulb brands (using large samples of both brands), a statistician used a normal distribution to find the *p*-value of the difference.



Which of the following statements ***best explains why*** the test for significance involves normal probabilities, rather than probabilities from the light bulb lifetime distribution?

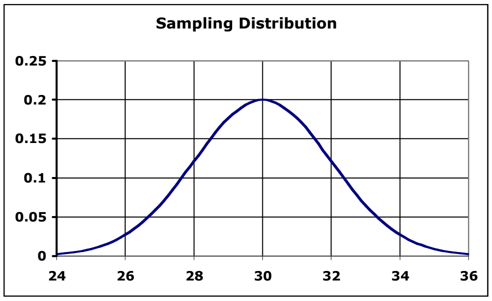
a. The distribution of the difference of means of large samples is always approximately normal.

b. The distribution of values in large samples is always approximately normal.

c. Values of the standard normal probability distribution are always given in reference tables.

d. The distribution of differences of values in two samples is always approximately normal.

9. Tests show that fuel efficiency of cars in the current model year averages 30 miles per gallon. A test of 100 new car models gave mean fuel efficiency of 31.5 miles per gallon. To see whether it is correct to claim that the new car models are more fuel-efficient than those in the current model year, a researcher constructed the sampling distribution of average fuel efficiencies of many samples of 100 current model cars shown in the following graph.



Which of these conclusions is ***best supported*** by the graph above?

a. The difference in fuel efficiency of current and new model cars is not statistically significant.

b. Half of current and new model cars have fuel-efficiencies below 30 miles per gallon.

c. The difference in fuel efficiency of current and new model cars is statistically significant.

d. Nothing can be concluded. The graph should be centered at 31.5.

10. When *Consumer Reports* studied response times for a random sample of 60 computer help-line calls, they found a mean of 15 minutes and standard deviation of 4.5 minutes. After hearing complaints about decline in service, they repeated the study (again using a sample of 60 calls) and found a mean response time of 16.5 minutes and standard deviation of 6.0 minutes.

What is the ***most plausible interpretation*** of the difference between the two study results?

a. Because the second study showed a higher mean, that study must have only looked at computer help-lines that received a lot of consumer complaints.

b. The increase in mean response time confirms a decline in services by computer help-lines.

c. The observed difference in mean response times is quite possibly due to chance variation.

d. The increase in standard deviation is the reason for the increase in mean response time.

~~~~~~

Teachers and school administrators are always interested in new textbooks or forms of instruction that will result in more success for their students. Items 11-14 ask about design, analysis, and interpretation of statistical studies to test the effectiveness of educational programs.
~~~~~~

11. To test the hypothesis that private schools are more effective than public schools, researchers plan to compare mean starting salaries of private and public school graduates. But they cannot agree on whether to test the results at a significance level of 0.10 ($\alpha = 0.10$) or at a significance level of 0.05 ($\alpha = 0.05$).

What effect will using 0.10 rather than 0.05 have on the study?

a. Using 0.10 will result in a greater chance that they will incorrectly retain the null hypothesis.

b. Using 0.10 will result in a greater chance that the null hypothesis is actually true.

c. Using 0.10 will result in a greater chance that they will incorrectly reject the null hypothesis.

d. Using 0.10 will result in a greater chance that the alternative hypothesis is actually true.

12. In an educational study the mean test score of students studying from a new, experimental textbook A was greater than that for students studying from a previously used, traditional textbook B, with significance at the $\alpha = 0.05$ level.

What action in response to that result ***makes most sense*** to you?

a. Compare the mean scores to see if the difference is great enough to merit the cost of new books.

b. Schools should adopt textbook A because its use leads to significantly better learning.

c. Re-analyze the data to see if the difference in means is significant at the 0.10 level.

d. Take no action until the study is repeated, because the difference in scores could be due to chance.

13. To test the effectiveness of a new method of teaching reading, researchers used the new method with a class of 35 second-grade students and found that 70% of those students were then reading above grade level. In a typical year, 50% of second-grade students are reading above grade level. In order to test the significance of the new program effect, researchers calculated the test statistic

$$z = \frac{0.7 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{35}}} \approx 2.37$$

What is the ***best explanation*** of what the researchers learn by using a statistical table to find a *p*-value for the test statistic 2.37?

a. The *p*-value tells the probability that the new teaching method results in a 20% gain in the number of students reading above grade level.

b. The *p*-value tells the probability that the new teaching method does not result in a 20% gain in the number of students reading above grade level.

c. The *p*-value tells the probability of getting the observed results, if the new program does result in better reading skill.

d. The *p*-value tells the probability of getting the observed results, if the new program does not result in better reading skill.

14. To evaluate a new computer-based approach to teaching pre-calculus, 200 volunteers among the 1000 pre-calculus students took the course on-line. At the end of the semester the mean final exam scores were 83.5 for the on-line students and 83.1 for the other students, a difference that proved to be significant at the 0.05 level.

If you were unhappy with the resulting recommendation that the course be taught on-line to all students, which of the following critiques is ***least likely*** to be effective in challenging the recommendation?

a. While the difference in means may be statistically significant, it is very small in practical terms.

b. The study did not use random assignment of subjects to treatment groups.

c. The experimental group was too small—only 200 out of 1000 students.

d. A previous experiment on the calculus course did not show positive results for on-line instruction.

APPENDIX C

FOLLOW-UP INTERVIEW QUESTIONS

Format: Semi-Structured, conversational

First Question: What is statistical hypothesis testing? If you had to explain it to someone, how would you explain what it is, what it does, how it's used?

Format for Remaining Questions: Look at item ________. Why did you choose the answer you chose and why didn't you choose the others?

Participants are told to explain their thinking as much as possible. In addition, they are told that they may change their answers, if they like.

The table below highlights important issues associated with each item that should be raised if they do not come up naturally in the course of the discussion.

Item	Code/Answer	Associated Issues to Raise
Method and Context		
1. Which of the following questions is ***most likely*** to be answered by a study that requires statistical hypothesis testing? a. Do athletes have a lower GPA than other students? b. What equation predicts a student's freshman GPA from his/her SAT score? c. What are typical costs for full-time resident students in U. S. colleges? d. Do the 12:00 noon sections of STAT 100 perform better than the 2:00 p.m. sections this semester?	RA 1.1, 1.2, 1.3 A	Consider similarities and differences of *a* and *d*. (*A* is the correct answer. Statistical hypothesis testing is not needed for *d* – there is no need to sample.)
10. When *Consumer Reports* studied response times for a random sample of 60 computer help-line calls, they found a mean of 15 minutes and standard deviation of 4.5 minutes. After hearing complaints about decline in service, they repeated the study (again using a sample of 60 calls) and found a mean response time of 16.5 minutes and standard deviation of 6.0 minutes. What is the ***most plausible interpretation*** of the difference between the two study results?	CS 1, 2 C	The question asks for the "most plausible interpretation". Is each option plausible, and only one is the best? Or, are one or more options implausible?

a. Because the second study showed a higher mean, that study must have only looked at computer help-lines that received a lot of consumer complaints. b. The increase in mean response time confirms a decline in services by computer help-lines. c. The observed difference in mean response times is quite possibly due to chance variation. d. The increase in standard deviation is the reason for the increase in mean response time.		
12. In an educational study the mean test score of students studying from a new, experimental textbook A was greater than that for students studying from a previously used, traditional textbook B, with significance at the 0.05 level. What action in response to that result ***makes most sense*** to you? a. Compare the mean scores to see if the difference is great enough to merit the cost of new books. b. Schools should adopt textbook A because its use leads to significantly better learning. c. Re-analyze the data to see if the difference in means is significant at the 0.10 level. d. Take no action until the study is repeated, because the difference in scores could be due to chance.	IP 1, 2, 3 A	This question asks what makes the *most sense*. Does each option make sense, but only one makes the most? Or, do one or more of the options make no sense?
Logic and Reasoning		
3. Which of the following statements is the ***best justification*** for using a statistical hypothesis test to answer the question: Are female students as successful as male students in college mathematics? a. It allows you to talk about uncertainty associated with your decision. b. It allows you to use only a sample of students to prove something for all students. c. It allows you to calculate means and use statistical tables to answer the question. d. It allows you to find and prove the answer using mathematical calculation.	RA 2 A	This question asks for the *best justification*. Is each option a justification, but only one is the best? Or, are one or more of the options not justifications?
4. A researcher would like to establish that freshmen in the humanities have higher SAT scores than freshmen in the sciences. Which of the following ***null hypotheses*** should be tested? The mean SAT score of humanities students is …	GH 1 C	Does the statement "the null is always equal" make sense? Could the alternative be the "equal one" and the null be the "not equal" one?

a. greater than that of science students. b. greater than or equal to that of science students. c. less than or equal to that of science students. d. less than that of science students.		
5. In 1950 the mean IQ of undergraduates at a university was 110. To test the hypothesis that students today are smarter, a study of 500 current students found a mean IQ of 120. The difference between the two means is significant at the 0.05 level. Which of the following statements is necessarily true? a. Undergraduates at the university today are smarter than those in 1950. b. The claim that undergraduates today are not smarter than those in 1950 is true with a probability less than 0.05. c. The claim that undergraduates today are smarter than those in 1950 has been established with 95% certainty. d. If undergraduates today are no smarter than those in 1950, the probability of the observed mean IQ is less than 0.05.	C 1, 2, 3 D	What are the null and alternative hypotheses? In your own words, what is each choice saying? Are there similarities and/or differences in the choices? What is the reasoning behind hypothesis testing and how do these options relate to that logic? How do the numbers cited in this item connect to the logic? (Connect back to definition of hypothesis testing the participant gave in the beginning of the interview) (As they talk about rejection regions, etc.) Did you use graphs when you did hypothesis testing? What do the graphs represent? What were the labels on the axes? (If they talk about tables, *z*-scores, etc.). What do the numbers in the table/*z*-score represent? Does this make sense to you?
6. A study tested the claim that: *Transfer students are less successful at the state university than students admitted as first time freshmen*. Results showed a difference in first semester grade point averages that is significant at the 0.05 level. Information from samples of transfer and first time freshmen is shown in the table below.	C 1, 2 C	The item asks for the *most reasonable inference*. All they all reasonable with only one the best? Or, are there one or more that are unreasonable?

	Transfer Admits	Freshman Admits
n	50	50
mean GPA	2.5	2.8

What is the ***most reasonable inference*** about the population of all first semester students that can be drawn from this information?

a. There are equal numbers of transfer and first time freshman students on campus.

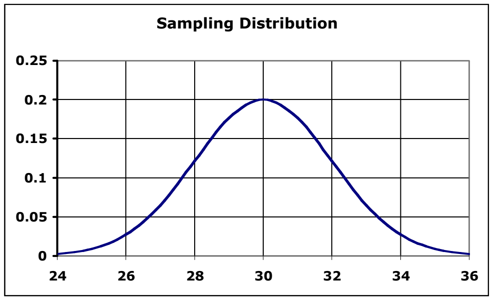
b. The mean first semester GPA of all freshman admits is 0.3 greater than that of all transfer admits.

c. It is unlikely that the first semester GPA of all transfer admits equals that of all freshman admits.

d. The mean first semester GPA of all University students is $\frac{2.5+2.8}{2}$ or about 2.65.

Be sure they understand what the table represents and that the difference was already found to be significant at the 0.05 level.

9. Tests show that fuel efficiency of cars in the current model year averages 30 miles per gallon. A test of 100 new car models gave mean fuel efficiency of 31.5 miles per gallon. To see whether it is correct to claim that the new car models are more fuel-efficient than those in the current model year, a researcher constructed the sampling distribution (for samples of size 100 of the current model) in the following graph.



Which of these conclusions is ***best supported*** by the graph above?

a. The difference in fuel efficiency of current and new model cars is not statistically significant.

b. Half of current and new model cars have fuel-efficiencies below 30 miles per gallon.

c. The difference in fuel efficiency of current and new model cars is statistically significant.

d. Nothing can be concluded. The graph should be centered at 31.5.

AS
2, 3

A

What are the null and alternative hypotheses in this situation?

What does the graph represent? What are the labels on the axes?
Did you use the graph in choosing your answer?

When you answered this, was there only one answer, or do you think several could be correct?

What do each of the choices mean? Can you explain them in your own words?

If possible, bring up the fact that the data supports the alternative. Doesn't that mean statistical significance automoatically?

11. To test the hypothesis that private schools are more effective than public schools, researchers plan to compare mean starting salaries of private and public school graduates. But they cannot agree on whether to test the results at a significance level of 0.10 ($\alpha = 0.10$) or at a significance level of 0.05 ($\alpha = 0.05$). What effect will using 0.10 rather than 0.05 have on the study? a. Using 0.10 will result in a greater chance that they will incorrectly retain the null hypothesis. b. Using 0.10 will result in a greater chance that the null hypothesis is actually true. c. Using 0.10 will result in a greater chance that they will incorrectly reject the null hypothesis. d. Using 0.10 will result in a greater chance that the alternative hypothesis is actually true.	DR 1 C	What are the null and alternative hypotheses? Note similarities and/or differences in the choices. Address rejection region, graphs, tables, *z*-scores issues here (from problem number 5) if not already addressed.

REFERENCES

American Statistical Association. (2005). *GAISE college report*. Retrieved December 12, 2006 from http://www.amstat.org/Education/gaise/GAISECollege.htm.

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

Bady, R. J. (1979). Students' understanding of the logic of hypothesis testing. *Journal of Research in Science Teaching, 16*(1), 61-65.

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education*, *38*(2), 115-131.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*(1&2), 75-97.

Bruce, C. (1991). *Exploring the sampling laboratory* (Educational Development Center, Inc. Center for Children and Technology Technical Report Issue No. 18). Retrieved October 26, 2006, from http://www.edc.org/CCT/ccthome/reports/tr18.html.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3). Retrieved January 19, 2005 from http://www.amstat.org/publications/jse/v10n3/chance.html.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In R. Lesh and A. Kelly (Eds.), *Handbook of Research Methodologies for Science and Mathematics Education* (pp. 341-385). Hillsdale, NJ: Lawrence Erlbaum.

Cobb, George. (1992). Teaching statistics. In Lynn A. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (MAA Notes No. 22), 3-43.

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

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*(2), 161-172.

delMas R. C. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, *10*(3). Retrieved January 26, 2005 from http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html delMas, R. C., Garfield, J., & Chance, B. (2004). *Using Assessment to Study the Development of Students' Reasoning about Sampling Distributions*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Evangelsta, F., & Hemenway, C. (2002). The use of the jigsaw in hypothesis testing. Second International Conference on the Teaching of Mathematics at the Undergraduate Level, Hersonissos, Crete, Greece.

Evans, J. St. B. T., & Newstead, S. E.(1995). Creating a psychology of reasoning: the contribution of Peter Wason. In S. E. Newstead, & J. St. B. T. Evans (Eds.), *Perspectivse on Thinking and Reasoning: Essays in Honour of Peter Wason* (pp. 1-16). East Sussex, UK: Lawrence Earlbaum Associates.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The psychology of deduction*. Hove, UK: Lawrence Earlbaum Associates Ltd.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83-96.

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

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. L. (2005). A curriculum framework for preK-12 statistics education. Presented to the American Statistical Association Board of Directors for Endorsement.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379-390.

Gal, I. (2004). Statistical literacy. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Statistical Literacy, Reasoning, and Thinking* (pp. 47-78). Dordrecht, The Netherlands: Kluwer.

Garfield, J. (1995). How students learn statistics. *International Statistics Review*, *63*(1), 25-34.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3). Retrieved January 19, 2005 from http://www.amstat.org/publications/jse/v10n3/garfield.html.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts: Implications for research. *Journal for Research in Mathematics Education*, *19*(1), 44-63.

Garfield, J., & Chance, B. (2000). Assessment in statistics education. *Mathematical Thinking and Learning*, *2*(1 & 2), 99-125.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education*, *10.* Retrieved December 12, 2006, from http://www.amstat.org/publications/jse/v10n2/garfield.html.

Gorman, M. E. (1995). Hypothesis testing. In S. E. Newstead, & J. St. B. T. Evans (Eds.), *Perspectives on Thinking and Reasoning: Essays in Honour of Peter Wason* (pp. 217-240). East Sussex, UK: Lawrence Earlbaum Associates.

Greene, J. C. (2001). Mixing social inquiry methodologies. In V. Richardson (Ed.), *Handbook of Research on Teaching* (pp. 251-258). Washington, DC: American Educational Research Association.

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65-97). New York: Macmillan Publishing Company.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge. In J. Hiebert (Ed.), *Conceptual and Procedural Knowledge: The Case of Mathematics* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hong, E., & O'Neil, H. F. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology, 82*(2), 150-159.

Hubbard, R., & Bayarri, M. J. (2003, August). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing. *The American Statistician*, *57*(3), 171-179.

Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neymen-Pearson views in textbooks. *Journal of Experimental Education*, *61*(4), 317-333.

Jacobs, V. R. (1999). How do students think about statistical sampling before instruction? *Mathematics Teaching in the Middle School*, *5*(4), 240-263.

Johnson, R. A., & Bhattacharyya, G. K. (2006). *Statistics: Principles and Methods* (5th ed.). New Jersey: John Wiley & Sons, Inc.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430-454.

Kilpatrick, J., Swafford, J., Findell, B. (Eds.). (2001). *Adding it up: Helping Children Learn Mathematics*. Washington, DC: National Academy Press.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development*, *71*(5), 1347-1366.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*(1), 59-98.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education*, *3*(1). Retrieved January, 26, 2005, from http://www.amstat.org/publications/jse/v3n1/konold.html.

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

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1995). Inconsistencies in students'reasoning about probability. *Journal for Research in Mathematics Education*, *24*(5), 392-414.

Krauss, S., & Wassner, C. (2002). *How significance tests should be presented to avoid the typical misinterpretations*. Paper presented at the Sixth Annual International Conference on Teaching Statistics, Cape Town, Africa.

Kuhn, D., & Dean, D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development*, *5*(2), 261-188.

Lajoie, S. P. (1998). Preface. In S. P. Lajoie (Ed.), *Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12* (pp. 3-32). New Jersey: Lawrence Erlbaum Associates.

Lane-Getaz, S. J. (2007). Development and validation of a research-based assessment: Reasoning about P-values and statistical significance. *Dissertation Abstracts International*, *68*(06). (UMI No. 3268997).

Link, C. W. (2002). *An examination of student mistakes in setting up hypothesis testing problems*. Louisiana-Mississippi Section of the Mathematical Association of America.

Lipson, K., Kokonis, S., & Francis, G. (2003). Investigation of students' experiences with a web-based computer simulation. *Proceedings of the 2003 IASE Satellite Conference on Statistics Education and the Internet*. Berlin. Retrieved October 18, 2007 from http://www.stat.auckland.ac.nz/~iase/publications/6/Lipson.pdf.

Liu, Y. (2005). Teachers' understandings of probability and statistical inference and their implications for professional development. *Doctoral Dissertation*, Vanderbilt University, Tennessee. Retrieved September 9, 2006, from http://www.stat.auckland.ac.nz/~iase/publications/dissertations/05.liu.Dissertation.pdf

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members'perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, *29* (4), 14-20.

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

Moshman, D. (2004). From inference to reasoning: The construction of rationality. *Thinking and Reasoning*, *10*(2), 221-239.

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

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339-363.

Pollatsek, A., Konold, C. E., Well, A. D., & Lima, S. D. (1984). Beliefs underlying random sampling. *Memory and Cognition*, *12*(4), 395-401.

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

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21-26.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3). Retrieved January 19, 2005 from http://www.amstat.org/publications/jse/v10n3/rumsey2.html

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

Scheaffer R. L., Watkins, A. E., & Landwehr, J. M. (1998). What every high-school graduate should know about statistics. In S. P. Lajoie (Ed.), *Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12* (pp. 3-32). New Jersey: Lawrence Erlbaum Associates.

Schwartz, D. L., Goldman, S. R., Vye, N. J., Barron, B. J., & The Cognition and Technology Group at Vanderbilt (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on Statistics: Learning, Teaching, and Assessment in Grades K-12* (pp. 233-274). New Jersey: Lawrence Erlbaum Associates.

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 Publishing Company.

Shaver, J. P. (1993). What significance testing is, and what it is not. *Journal of Experimental Education*, *61*(4), 293-316.

Snee, R. D. (1999). Discussion: Development and use of statistical thinking: A new era. *International Statistical Review*, *67*(3), 255-258.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*, 404-411.

Sternberg, R. J. (1999). Introduction to cognitive psychology. In R. J. Sternberg (Ed.), *Cognitive Psychology 2nd Edition* (pp. 1-26). Fort Worth, TX: Harcourt Brace College Publishers.

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

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105-110.

Tversky, A., & Kahneman, D. (1982). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 3-20). Cambridge: Cambridge University Press.

Tweney, R. D., & Chitwood, S. T. (1995). Scientific reasoning. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on Thinking and Reasoning: Essays in Honour of Peter Wason* (pp. 241-260). East Sussex, UK: Lawrence Earlbaum Associates.

Wason, P. C. (1967). On the failure to eliminate hypotheses: A second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking and Reasoning*. Harmondsworth: Penguin.

Watson, J. M., & Moritz, J. B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education*, *31*(1), 44-70.

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

Well, A. D., Pollatsek, A., Boyce, S. J. (1990). Understanding the effects of sample size on variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*, 289-312.

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

Wilkerson, M., & Olson, M. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology*, *131*(6), 627-631.