The Dissertation Committee for Jennifer Julia Kaplan Certifies that this is the approved version of the following dissertation:

Factors in Statistics Learning: Developing a Dispositional Attribution Model to Describe Differences in the Development of Statistical Proficiency.

Committee:

Phillip Uri Treisman, Supervisor

Martha K. Smith

Mary R. Parker

Susan B. Empson

Jonathan Jay Koehler

Factors in Statistics Learning: Developing a Dispositional Attribution Model to Describe Differences in the Development of Statistical Proficiency.

by

Jennifer Julia Kaplan, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2006

Acknowledgements

I would like to thank the faculty of Department of Mathematics at The University of Texas at Austin who believed in my abilities and potential from the beginning and who allowed me to complete the degree and program I wished to complete. My advisor, Uri Treisman was instrumental in making that happen. Without his support and advice, academic and personal, this project would not have been possible. I would also like to thank the members of my committee, Martha Smith, Mary Parker, Susan Empson and Jay Koehler, for their participation in my graduate studies. Martha Smith was always available to comment on my work and her contributions were invaluable. Mary Parker was influential in helping me become part of the national conversation in statistics education. Both Martha and Mary, along with Dr. Peter John, inspired and nurtured my appreciation for statistics as a discipline, providing my research with a direction.

I am grateful to my family and friends and who have supported me in all my endeavors. In particular, I want to thank my mom, who has read and edited this work both for content and grammar, and my dad, who would have, if he thought he would have understood it. Thanks also to Brett who has borne the brunt of my anxiety during the time I was working on my dissertation.

Special thanks go to Nancy Lamm. She always reminds me what I can do, and what I cannot. I am also grateful to Lilly, Susan and Carol for making sure I always had the time I needed.

Finally, I would like to recognize my former students and colleagues. When I think about my time teaching overseas, there are many names and faces that come to mind. They are too numerous to list here, but the students and faculty at the Istanbul International Community School and Colegio Internacional de Caracas know who they are.

Factors in Statistics Learning: Developing a Dispositional Attribution Model to Describe Differences in the Development of Statistical Proficiency.

Publication No._____

Jennifer Julia Kaplan, Ph.D.

The University of Texas at Austin, 2006

Supervisor: Philip Uri Treisman

This study sought to create a dispositional attribution model to describe differences in the development of statistical proficiency. The particular research question that guided the study was: To what extent can differences in psychological dispositions explain differences in the development of statistical proficiency and, in particular, students' understanding of hypothesis testing? In order to answer this question a framework to describe statistical proficiency was created. This framework guided the development of assessment materials used in this study. The results presented in this work are based on both large and small sample studies of undergraduate students who have taken an algebra-based statistics course. The large sample studies use quantitative methods to find relationships between statistics learning and dispositions. The small sample studies use qualitative methods from grounded theory to uncover themes and common conceptions and misconceptions held by undergraduate statistics students. No relationships were found between the statistics learning and the dispositions that were

studied, Need for Cognition and Epistemological Understanding. The research did identify several themes in the student discussions of hypothesis testing. The three emergent themes found were, how students consider the experimental design factors of a hypothesis test situation, what types of evidence students find convincing, and what students understand about p-values.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1. BACKGROUND

In her 1992 Presidential address to the annual meeting of the American Statistical Association, Katherine Wallman claimed that our society faces difficult problems, the solutions to which could benefit from contributions from statisticians. Some examples she cited were: maintaining a competitive advantage in the global marketplace, adapting to the changing composition of the workforce, overcoming pollution and other threats to the environment, and determining when to release new treatments for a disease (Wallman, 1993). She laments, however, that due to a

> series of "mis-es" – misunderstandings, misperceptions, mistrust, and misgivings: misunderstandings about the sources of statistical data; misperceptions of the willingness of citizens to provide information; mistrust by some members of our population about how statistical data will be used; misgiving about the value of statistics for guidance in public and private choices (pg. 1),

the benefit of statistics may not be brought to bear on important problems in an appropriate way. Wallman said that the "mis-es," which make it difficult collect and disseminate data in a useful manner, are "rooted in society's lack of 'statistical literacy'" (ibid).

Wallman further asserted that citizens of our society encounter statistics daily and do not have the necessary statistical understanding to evaluate the information. To address this problem, she asserted that trained statisticians should be deeply involved in the development of statistics and quantitative literacy programs for students of all ages. But what knowledge do statisticians need to have if they are to productively influence the development of students' statistical understanding? This work is designed to address the question by clarifying the meaning of statistical understanding and the various

psychological and dispositional factors that shape how students, and more broadly, citizens, make sense of and use statistical information.

In addition to its growing importance to citizenship, making sense of a world in which data plays a greater role, statistics is an increasingly important part of a growing number of college majors. A search of the course listings of the University of Texas at Austin for Spring semester 2003, for example, revealed undergraduate statistics and/or research courses in the departments of sociology, pharmacy, social work, psychology, and educational psychology as well as the department of mathematics. An initiative to formalize a division of statistics at the University counted upwards of twenty departments that offer quantitative analysis courses (Martha Smith, personal conversation, April 2003).

Nationally, enrollment in statistics among undergraduates and high school students has been growing rapidly. According to the 2006 College Report produced by the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project, in 1970, statistics enrollments were 27% of the size of calculus enrollments. By 2000, the percentage had risen to 74%. When the Advanced Placement (AP) statistics exam was first given in 1997, 7,500 students completed the exam. In 2004, the number of students completing the AP statistics exam was 65,000 (GAISE, 2006). Given the number of majors that require understanding of statistics and the growth of the population taking statistics classes, the study of statistical learning at the undergraduate level is increasing in importance.

The culminating idea of introductory statistics is inference: drawing conclusions about a population from sample data. Hypothesis testing is one aspect of statistical inference. Students in fields such as, pharmacy, nursing or psychology, are encouraged

to take statistics early in their academic career so they are prepared to read and interpret and later produce results based on hypothesis testing (personal conversations with students). There is anecdotal evidence to suggest that students can spontaneously make decisions about a population that is consistent with sample data, hence exhibiting intuitive understanding of the reasoning that underlies a hypothesis test. These same students, however, struggle to master the formal reasoning and enactment of a hypothesis test. In fact, Brewer claims that the area of inference is "the most misunderstood, confused, and abused of all … statistics topics" (1985, pg. 255).

> **An anecdote:** The professor enters the classroom with a new set of playing cards and a package of cookies. She tells them that they will play a game for the cookies. Each student will draw a card and anyone who draws a black card will receive a cookie. She breaks the seal on the cards and shuffles them and the students form a line to choose their cards. The first student draws a red card, as does the second. Red comes up third, fourth and fifth. At this point students begin to grumble amongst themselves. In fact, by the sixth red card, students at the end of the line begin to go back to their seats. At this point the instructor stops the activity and begins to discuss what happened. (Derived from personal communication, Roxy Peck, August 2004)

The instructor who told me this story used it as an illustration of her belief that individuals have the intuitive capability to perform the logical equivalent of a formal hypothesis test. She believes that the students at the back of the line are sitting down because they believe the class has been "duped." They think that there are no black cards and that no one is going to receive cookies. If they still believed that the deck was half black and half red, the normative behavior for the students at the end of the line would be to remain in line because every red card chosen early translates to a higher probability of a black card later. By sitting down, they affirm a belief that the deck is somehow stacked with reds. They do this without viewing the entire population; the judgment is based on a sample, just as it is in a formal hypothesis test. It is interesting to note, however, that at

the end of the narrative the instructor qualified the success of this simulation, admitting that students still have difficulty with formal statistical inference even after completing the activity. This serves as a reminder of the difficulty students have in linking their experiences with learning outcomes.

Three reasons given in the literature to explain the difficulties students have in learning the fundamental ideas of probability are 1) underlying difficulties with rational number concepts and proportional reasoning, 2) probability ideas are in conflict with the ways in which students view the world, and 3) students have developed a distaste for the subject as a result of having been exposed to it in an abstract and formal manner (Garfield and Ahlgren, 1988). It is the second of the three explanations that is of primary interest to understanding the background for the research presented in this paper. The idea that students bring to the class, or develop over time, misconceptions about the world that interfere with learning is not a new one. It has been studied by both physics and biology educations in a fair amount of depth (Garfield and Ahlgren, 1988, National Research Council, 1999). Research in mathematics has shown that students' "misconceptions of probability are particularly resistant to elimination" (Hirsch and O'Donnell, 2001).

An overview of the literature suggests using direct attempts to help students overcome their misconceptions, for example, by directly confronting students' statistical understanding or by limiting classroom examples (Garfield and Ahlgren, 1988). These efforts have limited success (Fischhoff, 1982). Instead of helping students to change their underlying intuitions, these interventions teach students rules by which to solve the problems (Sedlmeier, 1999 and delMas, et al., 1999). The research presented in this paper views misconceptions from a different standpoint. Parallel to the research on confronting subjects' misconception of stochastics is a body of research assessing

differences between subjects that account for differences in conception and misconception formation in reasoning (Stanovich, 1999 and Weinstock, and Cronin, 2003). Both bodies of literature will be discussed in more detail in the following chapters. The research presented here seeks to inform the second body of research, identifying dispositional factors that affect student learning in statistics, particularly in the area of hypothesis testing. In addition, this work sees to make relevant psychological research salient in the statistics education community.

The goal of this research is to identify differences in student learning that result from differences in dispositional factors among students. When one makes an assumption about the underlying conditions regarding the causes of the behavior, as is the goal of this work, he is said to be making an attribution (Corsini, 1999). When the attribution is based specifically on dispositional factors, the attribution is called a dispositional attribution (ibid). Disposition, as a word, has several meanings. The definition of disposition that is used in this work is that it is an underlying personality trait: something that is measurable and produces in a person the tendency to behave in a similar manner at different times and places (ibid).

The purpose behind the search for dispositions that affect statistics learning is not, of course, to be able to select future statisticians by examining the personalities of young children or college students. This work intends to inform statistics instructors of the different dispositions that may be affecting student learning so they can be sensitive to the different types of learners in their classroom. Further, this research seeks to model the progression of work done by Carol Dweck and her colleagues in the field of motivation theory. Dweck developed a research-based model of motivational processes in which it was shown that certain dispositions related to learning produce behaviors that are

maladaptive to productive learning (Dweck, 1986). The model was used as a basis for the creation of successful interventions that helped students with maladaptive dispositions develop dispositions more productive for learning (Dweck, 2002, Aronson, 2002). Once dispositions that describe productive statistics learners are found, research can be done to identify ways to develop those dispositions in all students.

2. RESEARCH GOALS

The primary goals of this research are to explore whether certain dispositional factors affect the development of statistical proficiency and, if so, to describe the magnitude of the effect. In order to achieve this goal, we must define, with care, the constructs we are examining and the theoretical constructs in which they operate. We begin by positing a definition of statistical proficiency. The framework that was designed for this study and its theoretical background are discussed in Chapter 2 of this work. Next, psychological constructs that might be part of the dispositional attribution model of statistical proficiency were identified. The basis for the selection of the constructs is developed in Chapter 3 of this work. Finally, instruments that could be used to assess the development of statistical proficiency were developed. These are described within Chapter 4, as is the description of the experimental design that was utilized in this research.

The specific question that is addressed by this research is: To what extent can differences in psychological dispositions explain differences in the development of statistical proficiency and, in particular, students' understanding of hypothesis testing? The population under study is those students taking a first semester algebra-based, rather than calculus-based, course in statistics.

3. METHODOLOGICAL ISSUES, LIMITATIONS, AND DELIMITATIONS

Study limitations stem from issues of internal validity while delimitations are the results of issues related to external validity, or generalizability (Dereshiwsky, 1999). It was not practical to use randomized design in this research. Therefore, this study will be largely observational in nature. This is a limitation in that it means that observed differences between subjects can not directly attributed to differences in dispositions.

The study does attempt to use large samples that represent the entire population where possible. Even in studies with large samples, the individuals tend not to be independent, attending the same university, having the same instructor, and/or using the same textbook. The research presented here will suffer from these problems. The results of the large-scale quantitative studies reflect a population that used the same textbook and had only three different instructors. A delimitation of this study, therefore, is that the findings may not be generalizable to other to other populations, even populations of undergraduate students at similar universities.

Due to limited resources and subject pools, it was not possible to use large samples in every aspect of this research. All of the studies that were completed as part of this work, however, were designed and funded for appropriate sample sizes. The use of volunteers and some aspects of the study design produced smaller samples than had been anticipated. The size of these samples limits the analysis that could be done and the scope of the conclusions that could be made from the data. It is important to utilize interview methods in order to find out what students are actually thinking (Garfield and Ahlgren, 1988). To adhere to proper design techniques, interview subjects were selected to meet several criteria based on the results of the larger scale studies rather than representing a simple random sample of students. This may reduce the variation in the backgrounds of

the subjects and counteract the limitation caused by the lack of randomization. On the other hand, it introduces a delimitation associated with selection bias.

One of the uses of interview techniques is to overcome the difficulty in finding out what students understand and how they will apply their knowledge outside of the classroom. On course assessments, students will give the response that they think the teacher wants, even when it does not match what they believe about a situation. Chinn and Samarapungavan (2001) studied this phenomenon using elementary school children's beliefs about water molecules. In order to uncover differences between what they call children's understanding and beliefs, the researchers asked their subjects to draw two pictures, one of what scientists think water looks like at a molecular level and another showing what they think water looks like at the molecular level (Chinn and Samarapungavan, 2001).

The research presented here takes several steps in an attempt to find out what students believe rather than what they think is the "correct" response. First, the investigator was in no way connected to the class. While the subjects knew they had been asked to participate in the study based on their having completed a statistics course, the were explicitly told that their instructor would not even know they had participated. On some instruments, subjects were specifically asked whether an answer they had provided was their true belief or the response they felt would receive high marks from their instructor. If those two types of responses would have been different from each other, the subjects were asked to record the differences.

In order to employ, efficiently, the opportunity provided by large lectures for large sample sizes that would strengthen the quantitative analyses, it was necessary to have subjects complete certain tasks during class time. When students complete a task

during class time, their identity as students may be salient. Therefore, caution may be needed in interpreting the results of tasks that were completed in-class. Rather than answers based on beliefs, students may have provided the answers they thought would be valuable or acceptable to the instructor. This bias may affect the internal validity of the study.

Finally, the use of qualitative analysis methodology introduces a limitation. Unlike most quantitative analysis methods, the same data, when analyzed qualitatively at two different times, might produce different results. Qualitative results lack internal consistency. This is usually overcome by the use of two or more independent analysts. The results can then be checked for inter-rater reliability. The research presented here is further limited by the fact that one researcher, with no corroboration, completed all qualitative analyses.

Chapter 2: Theoretical Framework to Describe Statistical Proficiency

1. INTRODUCTION

Over the last decade researchers within the statistics education community have recommended reforming statistics courses to help students develop statistical thinking, reasoning, and literacy skills rather than just the mastery of skills, procedures, and computations (Ben-Zvi and Garfield, 2004). In 1992, the ASA/MAA Joint Curriculum Committee gave as one of its three recommendations for changing statistics courses, incorporate more data and concepts and fewer recipes and derivations (Cobb, 1992). These initiatives coincide with the most recent efforts on the part of the National Council of Teachers of Mathematics to promote increased attention to the development of understanding in mathematics. Within the mathematics education community, the idea of increased focus on developing understanding is not a new one. Dewey and McLellen first called for increased emphasis on understanding in mathematics classes over 100 years ago in 1895 (Hiebert and Lefevre, 1986).

While the debate between placing emphasis on skill building versus understanding continues today within the domain of mathematics education, according to Hiebert and Lefevre (1986), the context for addressing the question of the relative importance of skills versus understanding has changed over the course of the last century. Previously, they claim, researchers wished to resolve the question of the relative importance of skill building versus understanding so they could create efficient and productive instructional programs for schools. By the mid-1980's, the goal of the research was to understand the process of knowledge acquisition and the relationship between types of knowledge as they are assimilated by the learner. The shift in focus

implies that an understanding of the process of learning and the interaction between knowledge types will provide a useful basis for building a coherent and effective instructional plan (Hiebert and Lefevre, 1986).

To this end, the National Research Council (NRC) has synthesized research from the cognitive sciences in order to describe how people learn. Their current publications include models for student learning in the domains of mathematics, history, and science. One particular publication, *Adding It Up: Helping Children Learn Mathematics*, draws on recent research in the cognitive sciences as a basis for creating a framework to describe proficiency in mathematics. Using similar sources, this chapter will suggest a description of proficiency in statistics in general. The framework will be used in future chapters to describe the assessment of proficiency in hypothesis testing.

Section Two of this chapter provides an overview of the model for proficiency, its component parts, shape, and a brief description of the literature on which it is based. In Section 3 the aspects of statistical proficiency are described. Included in the descriptions, which have been derived from literature in cognitive science and mathematics, science and statistics education, are the connections between the components and some examples of items that assess each strand. The chapter ends with a discussion of the weakness in the literature of one particular strand of proficiency, productive dispositions, and proposes an alternate definition of production dispositions in statistics learning.

2 OVERVIEW OF A MODEL TO DESCRIBE STATISTICAL PROFICIENCY

2.1 Components of the Model

The theoretical framework for statistical proficiency described in this chapter comprises five components: procedural fluency, conceptual understanding, strategic

competence, statistical reasoning, and productive dispositions. The heuristic use of the five question words, how, why, when, what and who, are used to briefly outline the differences between the aspects of competency. The question words are quite useful in identifying the main differences between the aspects of competency. Just as the distribution of a random variable varies about its center, however, so too do the aspects of proficiency vary around the central questions. If one considers each aspect of competency to be a distribution of a random variable, one understands that a complete description of each aspect of competency must include more information than just the location of the center. A more complete description of each aspect of competency is provided after the brief introduction.

- Procedural fluency – Learners know **how** to choose and complete appropriate processes
- Conceptual understanding – Learners know **why** the processes they use are appropriate
- Strategic Competence – Learners know **when** the use of statistical processes is relevant and can mediate processes and choices for the context in which they are operating
- Statistical Reasoning – Learners know **what** can be inferred from the analysis as well as **what** questions remain unanswered
- Productive Dispositions – Learners possess characteristics exhibited by people **who** are classified as expert statisticians

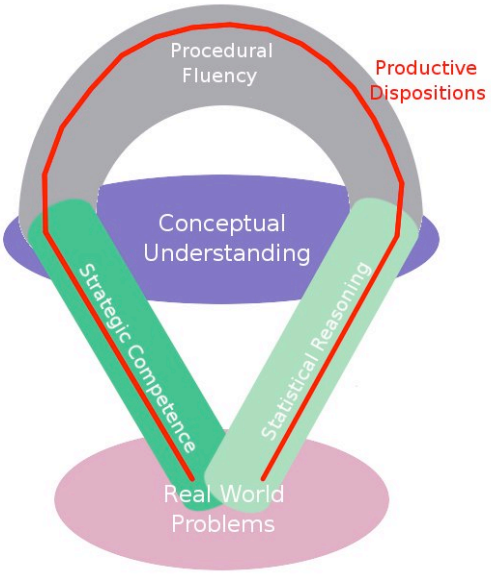
Procedural fluency is characterized by the ability to choose and complete appropriate processes, such as giving the correct summary statistics for a set of data. In the full description of conceptual understanding that appears later in this chapter, we will see that deep conceptual understanding is characterized in learners by the number of appropriate connections that exist between pieces of knowledge. More connections imply deeper understanding. Strategic competence is the aspect of proficiency that aids a learner in knowing when a statistical solution is appropriate. It also includes the ability to recognize structural features of statistical problems and to navigate appropriately between contextual information and statistical information. Statistical reasoning is the ability to make and critique an argument based on data. It includes knowing the rules for evidence within the domain of statistics. Productive dispositions are characteristics that help a learner to become a successful statistician. There is general agreement that dispositions impact student learning and growth of expert knowledge in subjects including statistics (see for example, Dweck, 1986, NRC, 1999, Wild and Pfannkuch, 1999). Some specific productive dispositions mentioned in the literature are self-efficacy in the domain, recognition of the value of the domain, and personality characteristics, such as, imagination and curiosity. The full description of productive dispositions in this chapter will suggest more robust constructs that might be used to measure dispositions that are productive to learning specifically in the domain of statistics.

2.2 Shape of the Model

The shape of the model for statistical proficiency is given in Figure 1.1. This figure describes statistical proficiency as it is manifested in an expert statistician. Notice that strategic competence and statistical reasoning create the connection between real

world problems and conceptual understanding. Furthermore, statistical reasoning and strategic competence stem from similar origins. Another feature of the model is the interaction of strategic competence, statistical reasoning and procedural fluency within

Figure 1.1: Full Model of Statistical Proficiency

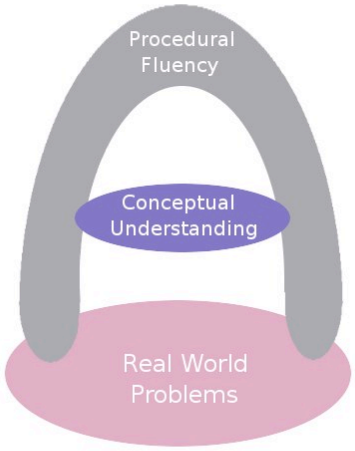


conceptual understanding. Notice that there is an area representing procedural fluency that lies outside of the area representing conceptual understanding. This indicates that there may be processes that statisticians can apply but that are not fully integrated into their web of understanding. Finally, productive dispositions may be thought of as a string or pump that helps the other aspects of proficiency to grow.

In contrast to Figure 1.1, Figure 1.2 provides a picture of the degenerate model of statistical proficiency that may be developed by a beginning student of statistics. In this model, strategic competence and statistical reasoning may not have developed at all.

Students must use their procedural fluency to solve real world problems. The small overlap between procedural fluency and real world problems indicates that the types of problems this student can solve are limited to those that are very similar to example problems posed by the teacher or textbook. Note also that procedural fluency in this model is physically thinner than in the fully developed model. Also, in this model, conceptual understanding is smaller and not yet well connected to procedural fluency.

Figure 1.2: Degenerate model of statistical proficiency



2.3 Theoretical Basis for Inclusion of the Strands

The first three components of statistical proficiency listed above, procedural fluency, conceptual understanding and strategic competence, form the portion of

statistical learning that has been called in the statistics education research "statistical thinking" (see for example Ben-Zvi and Garfield, 2004 and delMas, 2004). In creating the definitions of these competencies, I will be drawing on the research underlying the key findings that characterize the thinking and reasoning patterns exhibited by experts in their field. These findings are enumerated in the executive summary of *How People Learn: Brain, Mind, Experience and School* (NRC, 1999). The research will serve not only to differentiate between these cognitive skills, but will also provide insight into the relationships between procedural fluency, conceptual understanding, and the structure of knowledge that promotes a deep level of understanding. While the cognitive science literature will be used to describe the general ideas underlying the proficiency strands of strategic competence and productive dispositions, the more specific descriptions of these proficiency areas, as they pertain to doing statistics, will be taken from the literature in statistics education. The argument for the inclusion of statistical reasoning as a proficiency strand in the framework for statistical proficiency will be based on statistics education research. The description of this proficiency strand will be rooted in both the statistics and science education literature.

3. DETAILED DESCRIPTIONS OF THE COMPONENTS OF THE MODEL FOR STATISTICAL PROFICIENCY

3.1 Procedural Fluency

This work uses the same vocabulary, "procedural fluency" and "conceptual understanding," that is used by the NRC to describe the strands of mathematical proficiency (NRC, 2001). Within the cognitive science literature, the vocabulary used to

name these concepts differs from theory to theory (Hiebert and Lefevre, 1986). Piaget wrote about conceptual understanding and successful action; "Anderson distinguishes between declarative and procedural knowledge" (Hiebert and Lefevre, 1986, pg. 1). Perhaps the most rudimentary perspective on the difference between procedural fluency and conceptual understanding comes from a philosophical theory of learning in which the former is defined as "knowing how" and the latter as "knowing what" (Hiebert and Lefevre, 1986). Hiebert and Lefevre (1986) synthesize the ideas from the literature and define what they call "procedural knowledge", but which this work will call "procedural fluency," as having two parts. The first part is "composed of the formal language, or symbol representation system" (Hiebert and Lefevre, 1986, pg. 7) used by the domain. This suggests that successful learners in a domain should be familiar with "the individual symbols of the system and with the syntactic conventions for acceptable configuration of symbols" (ibid). The second type of procedural fluency consists of knowing the algorithms, rules and procedures for completing tasks within the domain. The NRC (2001) expands on this idea and proposes that procedural fluency encompasses all aspects of computational competence, from the choice of an appropriate process to being able to produce the correct computational result efficiently.

From this definition, one understands that at the heart of procedural fluency is the mastery of processes, which includes the ability to choose an appropriate procedure and complete it efficiently. There are many statistical processes for an expert statistician to master. Since, however, most of these processes are automated through the use of computer software, it may be argued that statisticians today do not need to be able to complete computations by hand using a process (Moore, 1990). Instead, procedural fluency in statistics is focused on knowing which process is appropriate in a given

context. In addition, procedural fluency in statistics requires knowledge of the symbols and syntax used by statisticians. One such example of syntax particular to statistics is the use of Greek characters to represent population parameters and Arabic characters to represent sample statistics.

3.2 Conceptual Understanding

Drawing from the cognitive science theories of learning, Hiebert and Lefevre (1986) define conceptual knowledge as follows:

> Conceptual knowledge is characterized most clearly as 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. Relationships pervade the individual facts and propositions so that all pieces of information are linked to some network. In fact, a unit of conceptual knowledge cannot be an isolated piece of information; by definition it is a part of conceptual knowledge only if the holder recognizes its relationship to other pieces of information.
>
> The development of conceptual knowledge is achieved by the construction of relationships between pieces of information. The linking process can occur between pieces of information that already have been stored in memory or between an existing piece of knowledge and one that is newly learned (pg 3 – 4).

This definition describes the structure of conceptual knowledge in which the connections between pieces of information are as important as the pieces of information themselves. Three of the key findings on expert learning include ideas about the organization and retrieval of information (NRC, 1999). They are,

- Experts have a great deal of content knowledge and the organization of the information reflects a deep understanding of the subject matter.
- Experts can retrieve knowledge with little attentional effort.

- Expert knowledge is conditional; it reflects the applicability of the knowledge and cannot be reduced to isolated facts or propositions.

The definition and findings above imply that as conceptual understanding deepens, not only are more facts contained within the system, but also, the connections between the facts adapt so as to create a more efficient retrieval system. Furthermore, an expert's web of knowledge is so well integrated that it can no longer be reduced to isolated propositions. Thus, the goal of developing conceptual understanding has as its foundation helping students make connections and develop rich webs of connected knowledge about statistical ideas.

Silver (1986) argues that while the analysis of and distinction between procedural and conceptual knowledge provided by Hiebert and Lefevre is thoughtful and insightful, it is "the *relationships* among, and not the distinctions between, elements of procedural and conceptual knowledge that ought to be of primary interest" (pg. 181). The basis of his argument is that distinctions are static, yet when one uses knowledge to perform a non-trivial task, the knowledge is used dynamically. Therefore, if we wish to understand learning beyond the solving of rote problems, we must consider the relationships between these types of knowledge. Hiebert and Wearne (1986) agree that an absence of connection between procedural and conceptual knowledge will lead to lack of competence. Furthermore, they claim that, at the time, the interaction between concepts and procedures was not well known (Hiebert and Wearne, 1986). While this may be the case, one can conduct a "thought experiment" about how procedural fluency and conceptual understanding interact within the domain of statistics.

Imagine the student who is asked to describe the distribution of a data set. The author of the question expects the student to provide a description of the shape of the data along with a computation of the measures of center and spread. The computation of the mean and standard deviation is not difficult given that there exist calculators that cost less than $20, such as the TI – 30 XIIS and the Casio FX-115MS Plus, which perform those functions. The calculation of the five number summary (the minimum, maximum, median and upper and lower quartiles) might be slightly more work on the part of a student if she does not have access to a graphing calculator. It is not, however, a difficult computation. What the calculator does not tell the student is the correct choice between the two sets of measurements of center and spread. So in this example, the primary object of procedural fluency being tested is the choice of the correct procedure; the efficient calculation of the measures is of secondary importance.

In order for a student to distinguish between the required measures, she must, at the very least, know that the five-number summary is appropriate to describe data that are skewed and the mean and standard deviation are appropriate to describe roughly symmetric data. At the very basic level, a piece of factual information may be used to distinguish the correct procedure. If the student has, as part of her conceptual understanding, the knowledge that the mean is affected by the presence of outliers while the median is not, she may envision that the tail of a skewed distribution could contain outliers, which would affect the calculation of the mean. Furthermore, she might have a deeper understanding of the mean as a balance point of the data which would allow her to think about the affect of the tail of the skewed distribution on the calculation of the mean even without the presence of outliers. The deeper her understanding of the mean, the less the decision between the two numeric summaries seems to be a memorized fact and the

more fluent the student will be in choosing measures of center and spread to describe data. At some point, the choice should become automatic, something that she need not think about consciously, as predicted by the studies of expert knowledge. This illustrates how a student may rely on conceptual understanding in order to choose the correct procedure, an action that is part of procedural fluency.

This example illuminates the relationship between procedural fluency and conceptual understanding in statistics proficiency at the most basic level. The stronger the structure of understanding becomes, the closer to expert will a learner come in exhibiting procedural fluency.

3.3 Strategic Competence

While students may believe that exhibiting procedural fluency and conceptual understanding on routine exercises demonstrates proficiency, clearly, experts do not complete tasks in their domain by a simple mechanism of listing all the facts pertinent to the problem and then applying the appropriate procedure (Resnick, 1987). For example,

> expert writers treat the process of composing an essay as a complex task of shaping a communication that will appeal to and convince an intended audience rather than as a simple task of writing down everything they know about a topic (pg 15).

Being able to choose and apply procedures and having a deep conceptual understanding of the ideas associated with a knowledge domain is not sufficient to demonstrate expert ability within that domain. The proficiency strand that describes what experts actually do as practitioners of their subject I will call “strategic competence.”

Strategic competence encompasses what an expert does when he is thinking and solving actual problems within his domain. Given that, it seems appropriate to incorporate research on expert problem solving to provide insight into this proficiency strand. One of the key findings on expert learning found by the NRC (1999) that applies to problem solving and has already been discussed in this chapter is:

- Expert knowledge is conditional; it reflects the applicability of the knowledge and cannot be reduced to isolated facts or propositions.

Two other key findings that are of importance to this part of the proficiency model are:

- Experts recognize meaningful patterns of information.
- Experts exhibit flexibility in their approaches to new situations within their domain of expertise.

These findings reflect flexibility and conditionality of knowledge. Experts use these facets of their understanding to apply what they know in new situations. Recognizing meaningful patterns of information aids experts in delineating the structural features of a problem, which we will see is a feature of expertise.

Meyer and Wittrock (1996) define problem solving as "a cognitive process directed at achieving a goal when no solution method is obvious to the problem solver" (pg. 47) and then go on to describe both the cognitive and metacognitive facets of the process. In terms of the cognitive factors that encourage development of problem solving, it has been learned that one striking difference between novices and experts in

many domains is the tendency of the novice to notice surface similarities, "such as characters and scenarios described in the problem" (NRC, 2001, pg. 125). In contrast, expert strategists focus on "structural relationships within problems, relationships that provide clues" (NRC, 2001, pg 125) that are useful in creating a representation or choosing a solution path.

The metacognitive processes for general problem solving, as described by Meyer and Wittrock (1996), are monitoring progress toward the desired goal and selecting and modifying the solution strategy as necessary. Furthermore, "advanced levels of problem solving are characterized by an increase in flexibility" (Carpenter, 1986, pg. 115) in choices of solution paths. Carpenter (1986) claims that the "increase in flexibility is made possible by an increasingly rich conceptual base, more efficient procedures and the maintenance of links between them" (pg 155). It is these connections that underlie the relationship between procedural fluency, conceptual understanding and strategic competence. Carpenter's conclusion might imply that as conceptual understanding and procedural fluency deepen, learners become more flexible in strategic competence and make gains along that strand of proficiency as well.

This interplay between the proficiency aspects of procedural fluency, conceptual understanding and strategic competence can be seen in the comparison between the degenerate form of the model of statistical proficiency and the fully developed form. In the degenerate form, procedural fluency is not well connected to conceptual understanding and only overlaps slightly with real world problems. The overlap with real world problems signifies the ability of beginning students only to solve problems similar to those posed in the text. As procedural fluency and its link to conceptual understanding grow, the strand of strategic competence also expands. Learners at this stage of

development of proficiency use strategic competence in conjunction with both procedural fluency and conceptual understanding to access more difficult real world problems.

I want to take the general ideas of strategic competence and make them specific to statistics. Wild & Pfannkuch (1999) propose a framework for statistical thinking in empirical enquiry that was developed through in-depth interviews with statistics students and professional statisticians. One dimension of their framework describes types of thinking. The authors discuss two types of thinking used by statisticians in the cycle of empirical enquiry, those that are general and those that are fundamental to statistical thinking. The following list is a subset of the types of thinking described in the framework that represent facets of strategic competence in statistics (Wild and Pfannkuch, 1999):

- Recognition of the need for data
- Consideration of variation
- Integrating the statistical with the contextual
- Transnumeration – creating meaning through representational changes
- Recognition and use of archetypes

One may consider strategic competence in statistics as the ability to think probabilistically about the world when a deterministic approach is not appropriate. Sophisticated statistical strategists have the ability to discern situations in which a probabilistic or statistical analysis is warranted in the decision-making process. The understanding of many of the "big ideas" and "types of thinking fundamental to statistical thinking" (Wild and Pfannkuch, 1999, pg. 226) are also included in this category. Using this description, the first two types of thinking listed address the component of strategic competence in which a statistical thinker recognizes that the need for a statistical

approach is appropriate. The last two types of thinking address the component of strategic competence in which the statistical thinker is able to discern the structural features of a situation that lead to an efficient solution. Integration of statistical and contextual knowledge bridges the gap between the two.

The recognition of the need for data is the core of strategic competence in statistics as it is the process by which one recognizes "the inadequacies of personal experiences and anecdotal evidence" (Wild and Pfannkuch, 1999, pg. 227) and turns instead to "deliberately collected data" (ibid) on which to base a conclusion. As noted previously, this tendency is closely related to reasoning because, when activated, it promotes statistical reasoning, reasoning that meets the standards for evidence and conclusions within the domain of statistics. The recognition of the need for data is, however, an aspect of strategic competence because the initial recognition, in a "real world" context, of the need for data rather than anecdote, is a strategic recognition. This recognition closely mirrors the notion of strategic competence in mathematics. Strategic competence in mathematics is defined by, and assessed through, the recognition that mathematics is the appropriate domain in which to investigate and make a decision (NRC, 2001). The recognition of the value of data over anecdote in statistics leads a learner to decide that a statistical solution is appropriate.

The consideration of variation similarly straddles proficiency strands. One can reason about the possible sources of variation and use a statistically crafted argument to explain or model the variation present in a situation. Many of the statistical procedures, such as analysis of variance, are processes that require a statistician to check assumptions about variance or that provide ideas about the nature of variance. The notion that variation is omnipresent underlies a deep understanding of concepts such as sampling

distributions. However, as the recognition of the need for data triggers the idea that statistical investigation is called for, so too does the consideration of variation trigger a statistical impulse. The consideration of variation is tied to the recognition of the need for data because it is the variation inherent in the world that allows a statistician to realize that an anecdote may be a story about a case that occurred as a result of sampling variation. Again, the impulse that leads to statistical reasoning is a manifestation of strategic competence that, as expertise grows, may be difficult to separate from proficiency in the strand of statistical reasoning.

Integrating the statistical ideas with the contextual features is ongoing during a statistical inquiry. It is the process by which a statistician decides how to appropriately use her prior knowledge of the context of the inquiry situation with her statistical knowledge and understanding (Wild and Pfannkuch, 1999). As the inquiry progresses, she will move between the spheres of contextual and statistical knowledge, choosing processes, revising hypotheses and forming conclusions in a cycle of inquiry. Furthermore, this integration skill will interact with a statistical disposition, described later in more detail, the openness to ideas that challenge preconceptions. A statistician must view data in context, but must also be open to data that are in conflict with prior beliefs. Hence, negotiating between context and data represents a strategic competence within the domain.

Transnumeration is a word coined by Wild and Pfannkuch (1999) to describe the "process of changing representations to engender understanding" (pg 227) that occurs when a statistician finds ways to capture meaning and facilitate understanding from the data. Creating appropriate representations, or what in statistics is often finding appropriate models, is, within the mathematical framework, an aspect of strategic

competency (NRC, 2001). Since transnumeration is such an activity, it is appropriately assigned to this strand of statistical proficiency. Finally, recognition and use of archetypes comprise the process through which statisticians recognize the applicability of a process based on problem type. This is a straightforward application of the aspect of strategic competence in which a proficient statistician uses structural features of a problem to categorize it and find an appropriate solution.

3.4 Statistical Reasoning

Recall that statistical reasoning is one of the specific areas mentioned by statistics educators as being in need of more emphasis. These researchers claim that the term "statistical reasoning" is not well defined in the literature and may be used to mean different things by different writers (Ben-Zvi and Garfield, 2004). Furthermore, definitions for statistical reasoning and statistical thinking provided in the literature tend to overlap and make it difficult to distinguish, from the definitions, the difference between the two activities (delMas, 2004).

While delMas acknowledges that both activities, reasoning and thinking, may be occurring in concert when one is doing statistics, he also proposes distinct definitions for the two ideas. He claims that "a person who knows when and how to apply statistical knowledge and procedures" (pg. 85) is demonstrating proficiency in statistical thinking. "By contrast, a person who can explain why results were produced or why a conclusion is justified demonstrates statistical reasoning" (delMas, 2004, pg. 85). The delMas definition of statistical reasoning still fails to distinguish reasoning from the other aspects of statistical proficiency. The ability to explain why something is true or appropriate has already been classified as a facet of conceptual understanding. Furthermore, delMas

seems to be thinking of reasoning as "making an argument" rather than taking a more broad view of statistical reasoning.

A search of the ARTIST Assessment Database maintained by the research group of which delMas is a part further confuses the distinction between statistical thinking, reasoning, and literacy. The following three items resulted from a database search for assessment items targeting hypothesis testing:

ARTIST item designed to measure **statistical thinking** within the domain of hypothesis testing:

> A medical researcher wishes to investigate the effectiveness of exercise versus diet in losing weight. A group of 25 overweight adult subjects have two different treatments. First, they are placed on a regular program of vigorous exercise, but with no restriction on diet. Then they are placed on a strict diet, but with no requirement to exercise. The weight losses after 20 weeks on each treatment are determined and the differences in weight loss between two treatments are computed for each individual. The mean of these differences in weight loss is found to be -2 lb. with standard deviation s = 4 lb. Is this evidence of a significant difference in mean weight loss for the two methods? Please justify.

ARTIST item designed to measure **statistical literacy** within the domain of hypothesis testing:

> The company Mason-Dixon Polling and Research, Inc. conducted an opinion poll for the St. Paul Pioneer Press and Minnesota Public Radio from Oct. 30 through Nov. 1, 2002, just prior to the election on Nov. 5, 2002. The poll surveyed 625 potential voters. One of the questions asked if the person thought Walter Mondale was the best choice to replace Paul Wellstone as the Democratic candidate for Minnesota senator. Of the 625 people, 344 answered YES to this question. Does the evidence from this sample support the hypothesis that more than half of all potential voters thought Mondale was the best choice? Justify your answer with an appropriate statistical procedure.

ARTIST item designed to measure **statistical reasoning** within the domain of hypothesis testing:

> A nutritionist thinks the average person with an income below the poverty level gets less than the recommended daily allowance (RDA) of 800 mg of calcium. To test her conjecture, she obtains the daily intakes of calcium for a random sample of 45 people with incomes below the poverty level. The mean of the sample is 737.3 mg with sample standard deviation of 262.2 mg. Is there sufficient evidence at the alpha = 0.05 level to support the researcher's claim?

Each of the three examples appears to ask the same question. The student must perform a hypothesis test in order to report whether the difference is statistically significant. The only difference among the questions is in the amount of guidance the student is given. In the "thinking" item, the student must realize that she is meant to perform a hypothesis test. In the "literacy" item, the student is cued to use a statistical procedure. The "reasoning" item is the most specific of the three, giving the student an alpha level at which to make the judgment. Within the framework being developed in this chapter, all of the questions above would be considered assessments of procedural fluency, testing whether the student could perform a hypothesis test and write an appropriate conclusion.

While it seems that we will have to work a bit harder to define statistical reasoning, delMas' definition of statistical thinking indicates that the proficiencies of procedural fluency, conceptual understanding, and strategic competence together combine to form the basis for "statistical thinking," This notion is consistent with delMas as he invokes the ideas presented by Wild and Pfannkuch (1999) and used here to describe strategic competence as the basis for his definition of statistical thinking (delMas, 2004). Furthermore, the question above, designed to test statistical thinking is the most general of the questions. While it may not tap into conceptual understanding per

se, the question is broad enough, not specifying how the response should be justified, that the area of strategic competence responsible for activating the knowledge that a statistical process is necessary must be accessed by the learner.

In order to deepen the definition of statistical reasoning, one may import ideas about reasoning from science education. According to delMas, (2004, pg 84):

> In recent years, statisticians have pointed out the distinctions between statistics and mathematics in order to establish statistics as a separate and unique discipline. Statistics may be viewed as similar to disciplines such as physics that utilize mathematics, yet have developed methods and concepts that set it apart from mathematical enquiry.

The assertion made be delMas links statistics to science in a way that suggests that applying ideas from science education to describe statistical proficiency is reasonable. The following passage, which was authored by a physics instructor discussing what he hoped his students would learn through the use of guided inquiry techniques in his classroom, further illustrates the connection between science learning and statistics:

> I want my students to understand and be able to apply the concept[s]…in real life situations…I also want my students to understand the nature of scientific practice. They should be able to interpret and explain common phenomena and design simple experiments to test their ideas. In short, I want them to have the skills necessary to inquire about the world around them, to ask and answer their own questions, and to know what questions they need to ask themselves in the process of thinking about a problematic situation (Minstrell and Krauss, 2005, pg, 476).

It seems that, if one were to exchange the phrases "statistical inquiry" for "scientific practice," "statistical results" for "common phenomena" and "research situation" for "problematic situation," the above paragraph could have been written by a statistics instructor discussing what she hopes her students would learn upon successful completion of a statistics course.

Consider the following passage taken from recent writing in statistics education about the goal of teaching statistical thinking:

> STATISTICAL THINKING involves an understanding of why and how statistical investigations are conducted. This includes recognizing and understanding the entire investigative process (from question posing to data collection to choosing analyses to testing assumptions, etc.), understanding how models are used to simulate random phenomena, how data are produced to estimate probabilities, recognition of how, when, and why existing inferential tools can be used, and being able to understand and utilize the context of a problem to plan investigations and draw conclusions. (ARTIST)

Both the passage authored by the science instructor and the passage authored by the statistics educators express the expectation that successful students will learn to complete the process of inquiry within the domain specified. Furthermore, both passages assert that successful students should understand the need for such a process as part of understanding the world in which they live. The similarities between these two passages demonstrate the close relationship between ideas surrounding proficiency in science and proficiency in statistics. This relationship suggests that the science education literature is a reasonable domain from which to import ideas about learning into the domain of statistics.

The National Science Education Standards (NSES) includes, in its definition of scientific literacy, the ability to "engage in social conversation about the validity of the conclusions" (NRC, 1996, pg 22) of scientific studies as reported in the media. Additionally, the NSES expects that a scientifically literate citizen,

> should be able to evaluate the quality of scientific information on the basis of its source and the methods used to generate it. Scientific literacy also implies the capacity to pose and evaluate arguments based on evidence and to apply conclusions from such arguments appropriately (ibid).

According to the editors of Learning and Instruction: A SERP Research Agenda, researchers, including Bazerman, Lemke and Kuhn have "pointed out that science entails mastering and participating in a particular form of argument (NRC, 2003, pg. 132)." Thus, reasoning, from the perspective of science literacy, incorporates the ideas of evidence and being able to judge the quality of evidence and the validity of arguments and conclusions using communally understood standards for evidence.

Incorporating the ideas about reasoning in science, I propose a definition for statistical reasoning that includes the ability to make statistically valid conclusions, critique statistical arguments, and discuss the scope of the conclusions generated. Reasoning is about knowing what can be inferred and what questions remain unanswered. Perhaps we may rewrite delMas' definition of statistical reasoning in this way: a person who knows what can be inferred from the data or results and whether a conclusion is valid demonstrates statistical reasoning.

After this reformulation of the definition of statistical reasoning, it is clear that reasoning capabilities are closely linked with strategic competence, which triggers the need for a statistical argument in the mind of a statistician. This is illustrated by its placement in the model next to strategic competence, connecting real world problems to the other aspects of proficiency. Central to the action of generating statistical arguments are the activities of reasoning from data and using statistical models (Wild and Pfannkuch, 1999). Other aspects of statistical reasoning involve understanding the scope of a conclusion. This includes, for example, distinguishing between association and causation when investigating the relationship between two or more variables. The ability to discuss correctly the scope of a statistical conclusion is intertwined with conceptual understanding. Most statistical conclusions are probabilistic in nature so an

understanding of probability as applied to the situation assists the statistician in differentiating between subtly different statements of conclusion. Garfield (2003) supports both the definition of statistical reasoning presented here and the reliance of the development of this competency on the development of conceptual understanding.

3.5 Productive Dispositions

The final strand of the proposed model for proficiency in the domain of statistics is that of productive dispositions. Learning theorists have shown that there are dispositions, habits of mind and ways of thinking about the world, which affect learning across domains. In general, positive attitude towards a subject, including the view that a subject is useful and worthwhile and a belief in one's abilities to learn a subject, have been shown to be productive to the learning endeavor (Schoenfeld, 1992). Self-concepts that are useful to success in developing proficiency in mathematics in particular include a high sense of self-efficacy and a belief in the value of diligence (National Research Council, 2001).

Motivation theory is one field in which researchers have identified dispositions that affect learning across domains. Through empirical studies with children, Carol Dweck and her colleagues have identified adaptive and maladaptive motivational patterns, where an adaptive pattern promotes learning and a maladaptive pattern inhibits learning (see, for example, Dweck, 1986, 2002, and Dweck and Leggett, 1988). Dweck has identified two theories of intelligence held by students (Dweck, 2002). Some students tend to believe that intelligence is fixed, that each person has a pre-determined, limited capacity to learn. Other students tend to believe intelligence is malleable that what one knows and can do may be improved through practice.

Dweck has shown that children who believe that intelligence is malleable tend to develop motivational patterns that are adaptive to learning. Children with this belief tend to have, as their educational goal, a desire to increase their own level of competence. Whatever their belief in their current ability, these learners tend to seek challenging tasks and to persist even when faced with failure. They view failure not as a limit on what they are capable of doing in the long term but as an indication of the boundaries of their current capabilities. They can use each failure as a learning tool that helps them to understand and to overcome their current deficiencies in competence. This pattern is adaptive in that the learner chooses difficult tasks from which she can learn and she persists even after failure. Therefore, the behavior leads to successful learning.

In contrast, students who view intelligence as fixed want to show what they know; they have a performance goal for learning activities. Those who believe their capacity for learning in a domain is high tend to select challenging tasks in order to show the depth of their knowledge. Those who believe their capacity for learning in a domain is low will avoid challenges in order to hide what they do not know. This belief pattern is maladaptive for learning because failures appear to lower a learner's belief about her abilities, which may cause her to begin to shy away from challenging tasks. She would no longer be giving herself opportunities to learn and expand her knowledge.

In addition to the dispositions that encourage learning across domains, the socio-constructivist school of learning theory proposes that, within any domain, there is a culture of practice that is assimilated by practitioners as they become experts in the field (Schoenfeld, 1992). This culture includes "a way of thinking, a way of seeing, and having a set of values and perspectives (Schoenfeld, 1992, pg. 340)." Some of the ways that statisticians think and view the world, for example, the consideration of variation and

the understanding of the need for data, have already been discussed under the heading of strategic competence. In their classification of statistical thinking as evidenced by statisticians, Wild and Pfannkuch (1999) also enumerate dispositions, what they also call "personal qualities (pg. 233)", exhibited by professional statisticians.

- Diligence
- Awareness and Curiosity
- Skepticism
- Openness to ideas that challenge preconception
- A propensity to seek a deeper meaning
- Being logical
- Imagination
- Engagement

As in any domain, diligence is a quality that aids a statistician or statistics learner to master the intricacies of the domain. Awareness and curiosity leads the statistician to notice when events are unusual or interesting, thereby beginning the process of statistical inquiry. She may see an unusual datum reported in the news or hear a story that defies her interpretation of the world. Next her sense of curiosity leads her to ask why this event has occurred, whether it is a general phenomenon or how it can be exploited. The attribute of curiosity is so crucial to the inquiry cycle that Wild and Pfannkuch (1999, pg. 233) explicitly wonder whether it is at the heart of all scientific inquiry and not limited to statistical inquiry.

Skepticism is the propensity to question the validity of reported facts or results once they have been noticed through the activation of the awareness disposition. It is the tendency that statisticians exhibit to ask whether a conclusion is indeed warranted by the

data or if the data are even plausible. It may be characterized by the tendency to be "constantly on the lookout for logical and factual flaws when receiving new ideas and information" (Wild & Pfannkuch, 1999, pg. 234). Openness to ideas that challenge preconceptions may be considered a subcategory of skepticism in that it requires the statistician to doubt, set aside, or be skeptical of her initial intuitions or preconceived opinions of the event in question. The psychology literature on human reasoning will show that subjects are less willing to believe logical arguments that lead to a conclusion they do not believe and more willing to believe illogical arguments that lead to a conclusion which they believe (see, for example, Evans, Barston and Pollard, 1983). Thus, the disposition of challenging one's preconceptions is not common amongst the general population, and its development would aid in a student's development of statistical proficiency.

The propensity to seek a deeper meaning is the disposition that keeps a statistician working on a problem, trying all avenues of exploration the way an investigator follows up leads until she is sure that she has found the truth of an event. Without this disposition, a statistician might notice a curious statement and be skeptical of its veracity, but she would be able to leave it at that, a curiosity that may or may not be a truth. The propensity to seek a deeper meaning is the personality trait that drives the statistician to create a representation with which to uncover the truth about the situation.

While a lay person might believe that logic and imagination as dispositions are contradictions of each other, in fact, the two moderate each other and both are necessary for a statistician to do her job. One needs imagination to develop representations for new problem situations or to generate explanations and models for data. Often it is a new representation that maps what had previously seemed to be an intractable problem onto a

representation that creates an easily solvable problem. Logic, "the ability to detect when one idea follows another and when it does not (Wild and Pfannkuch, pg. 234)" keeps one's imagination from running away unbridled. It allows the statistician to assess the representations developed by the imagination for logical construction and robustness. Furthermore, it prevents her from disregarding important cases, making fallacious arguments or providing illogical or unfounded explanations for features of her data.

While all of the dispositions discussed contribute to a statistician's proficiency and likelihood of arriving at a conclusion, according to Wild and Pfannkuch (1999), it is the disposition of engagement on which the development of the other dispositions relies, for any given situation. As a statistician becomes more interested in, or engaged with, her problem, she develops an "awareness…towards information on the peripheries of" (Wild and Pfannkuch, 1999, pg. 233) her experience that are related to the problem, but which she might have overlooked had she not been so occupied with the problem. Furthermore, an interest in the problem keeps the statistician involved in the statistical inquiry cycle until she is no longer curious about the situation. As engagement increases, the level of explanation that is required to satisfy her curiosity also increases. Wild and Pfannkuch (1999) suggest that students suffer from a lack of engagement because the problems with which they are presented are not interesting enough. Students of statistics would benefit from explicit attention by educators to the development of this disposition, as they would from attention to the disposition of openness to ideas that challenge preconceptions.

Within the discussion of each of the dispositions, one finds again the connections between the strands of proficiency. Skepticism as a disposition triggers the need for data, and the propensity to seek a deeper meaning suggests the need for new representations,

both facets of strategic competence. Imagination helps a statistician to create new representations, thus strengthening her strategic competence. In general, the curiosity disposition drives a statistician to learn more about the methods of reasoning from data and procedures by which one may learn about the world, enhancing statistical reasoning and procedural fluency. Within the shape of the model that describes statistical proficiency, productive dispositions are like pump that helps to inflate the other aspects of proficiency.

4. AN ALTERNATE DEFINITION OF DISPOSITION

According to The Dictionary of Psychology (Corsini, 1999) a disposition is a tendency to behave in a similar manner at different times and places. It has as a synonym the word trait, which is defined by the same volume to mean an enduring personality characteristic that determines a person's behavior. The key word in the definition of TRAIT is "enduring"; the difference between a personality trait and a personality type is that a trait is measurable and enduring and a type is a single term description, such as brilliant, stingy, cruel, lazy, and sweet. In general, personality traits can be measured by instruments that have been validated using factor analytic studies (Corsini, 1999); personality types are descriptors and are not measured. The Wild and Pfannkuch description of dispositions was based on the results of interviews with statisticians not on the outcomes of reliable instruments. Therefore, from the point of view of psychology, the dispositions they described are better classified as personality types. As part of this research, I hope to identify dispositions that are personality traits and productive to the learning of statistics. The remainder of the chapter describes two dispositions that, it will be argued, may be dispositions that are productive to statistics learning.

4.1 Need for Cognition

According to Cacioppo and Petty (1982), Cohen and his colleagues first discussed the construct, later termed the "Need for Cognition" (NC), in 1955. They described the attribute as a tendency for a person to want to "structure relevant situations in meaningful, integrated ways" and "understand and make reasonable the experiential world" (Cacioppo and Petty, 1982, pg. 121). People with a high need for cognition find it fun to think and may be perceived to be on a quest for reality (Cacioppo and Petty, 1982). The need for cognition is distinguishable, they claim, from cognitive style (Cacioppo and Petty, 1982). NC describes an individual's propensity to elaborate on events and to uncover reality, but does not specify the manner in which that is done. Cognitive style is defined as the characteristic way a person thinks about problems and conceives and implements solutions (Corsini, 1999).

Cacioppo and Petty (1982) developed the original scale to assess a subject's NC using the two populations, college faculty (high NC group) and industrial factory workers (low NC group). Subjects responded to 45 questions such as "I really enjoy a task that involves coming up with new solutions to problems" using a –4 to +4 Likert type rating scale. Thirty-four of the forty-five items discriminated between the faculty and the workers at $\alpha = .1$ under an ANOVA procedure. The responses to each of the subsets of discriminatory items was highly correlated to the total score accrued by subjects on the subset indicating an internal consistency within the subset of discriminatory items. The result of a principle components analysis indicated that only one factor was needed to describe the NC construct. In 1984, Cacioppo, Petty and Kao published a more efficient version of the NC assessment that contained a subset of 18 of the original 34 items.

Responses to the short version also produced one dominant factor, the results were highly correlated with the results of the long form and the reliability was nearly equivalent when Chronbach's alpha was computed. Therefore, it has been the standard to use the 18 item short assessment for NC since that time.

In a more recent publication, Cacioppo, et al. (1996) claim that all individuals, both those high and low in NC "must make sense of their world (pg. 198)." The difference between those high and low in NC is in the way that the individuals come to understand the world in which they live. Those high in NC "tend to seek, acquire, think about, and reflect back on information to make sense of stimuli, relationships, and events in their world (ibid)." In contrast, those low in NC "rely on others (e.g., celebrities and experts), cognitive heuristics, or social comparison processes (ibid)" to make judgments and provide structure for understanding the world. The theoretical definition of the construct of NC seems to describe what Wild and Pfannkuch (1999) claim statisticians are doing during the cycle of empirical enquiry, reasoning from data rather than anecdotes and using of analysis rather than heuristics in decision making.

Empirically, NC has been shown to be positively correlated with basing judgments and beliefs on empirical information and rational considerations, seeking out and scrutinizing relevant information in problem solving, being curious and open to new ideas (Cacioppo, et. al., 1996). These positive correlates to NC form much of the basis for Wild and Pfannkuch's (1999) description of personality type exhibited by expert statisticians. Both the theoretical description and the empirical correlates of Need for Cognition suggest that the construct may be a reasonable measure of personality traits that are specific to the domain of statistics as discussed in the model for statistical proficiency.

4.2 Epistemological Understanding

Epistemological Understanding is a construct designed to describe people's beliefs about the nature of knowledge and what it means to "know." It was developed by Deanna Kuhn and her research group at Columbia University and has been used in conjunction with juror reasoning studies. The theory was developed in the tradition of Perry, who studied the ethical development of undergraduate students in the late 60's, except that Kuhn's model accounts for development from childhood to adulthood and for subjects understanding of "knowledge" across domains of judgment whereas Perry studied, specifically, undergraduates' ethical development.

Kuhn's model for epistemological understanding has four levels, Realist, Absolutist, Multiplist and Evaluativist, each described by a subject's belief about assertions, reality, knowledge and critical thinking. While there are several models to describe epistemic reasoning or epistemological understanding, they differ in vocabulary, the words that are used to name the stages, but not in substance (Krettenauer, 2005). Only very young children are realists (Kuhn, Cheney and Weinstock, 2000); they view assertions as copies of an external reality that is directly knowable. Therefore, knowledge stems from an external source and is certain and critical thinking is unnecessary. As children move into an Absolutist phase, they view assertions as facts that may be correct or incorrect in their representation of reality. Their ideas about reality and knowledge have not changed, but they now view critical thinking as a vehicle for comparing assertions to reality in order to determine the truth-value of an assertion.

By upper elementary school, most subjects have made the transition to the Multiplist stage (Kuhn, Cheney and Weinstock, 2000) in which assertions are

characterized as opinions chosen by those who hold them and only accountable to the holder. For these people, reality is not directly knowable as knowledge is uncertain and generated by human minds. Thus, the use of critical thinking to discriminate the truth-value of an assertion is irrelevant. Many adults remain in the Multiplist stage and do not progress further. Some people, however, do attain the Evaluativist level of epistemological understanding. This level is characterized by the belief that assertions are judgments that can be evaluated based on standards of argument and evidence. Their views on reality and knowledge match those held by the Multiplists, but critical thinking

Table 1.1: Description of Epistemological Understanding Stages

	Assertions	Reality	Knowledge	Critical Thinking
Realist	COPIES that represent an external reality.	Directly knowable	External and certain	Unnecessary
Absolutist	FACTS that are correct or incorrect in their representations of reality (possibility of false belief).	Directly knowable	External and certain	A vehicle for comparing assertions to reality and determining their truth or falsehood
Multiplist	OPINIONS freely chosen by and accountable only to their owners.	Not directly knowable	Human generated and uncertain	Irrelevant
Evaluativist	JUDGEMENTS that can be evaluated and compared according to criteria of argument and evidence.	Not directly knowable	Human generated and uncertain	Valued as a vehicle that promotes sound assertions and enhances understanding.

is now a vehicle that promotes sound assertions and enhances understanding. The descriptions are encapsulated in Table 1.1, which was originally published in Kuhn, Cheney and Weinstock, (2000, pg. 311).

Consider the Evaluativist. She believes that assertions are judgments based on argument and evidence with the reality unknowable. This sentence could have been written about the statistician. The statistician does not know the reality of a population from which she is sampling; she uses evidence and a probabilistic argument to make a judgment about the population. The understanding of a statistical argument, or proficiency in statistical reasoning, may, thus, require a learner of such to hold an Evaluativist belief about knowledge. Similarly, a Multiplist learner may struggle to understand the reasoning behind hypothesis testing. She may find the task irrelevant since knowledge is an opinion and proof of its veracity is implicit in the personal conviction of the holder of the opinion. Perhaps a Multiplist outlook is a mechanism behind the value people place on anecdotal evidence over statistical evidence. As most subjects exhibit at least a Multiplist view from late adolescence onward, the population of students in an introduction to statistics class should be categorized as either Multiplist or Evaluativist. The above discussion suggests that the Evaluativists will exhibit deeper proficiency in the statistical reasoning component of hypothesis testing.

In the early studies by Kuhn and her colleagues, epistemological understanding was assessed through the use of the Livian War Task described by Weinstock and Cronin (2003). In this task, subjects read and heard two accounts of the fictitious Fifth Livian War, a civil war between North and South Livia. An historian from North Livia "wrote" one account and an historian from South Livia "wrote" the other. As a result, the two accounts conflict with each other. Using interviews, subjects were asked about the nature and sources of the differences between the two accounts, including questions about whether both accounts could be correct and whether there was any way to know for certain what had happened in the war. Furthermore, subjects were asked whether a third

account of the war might be different from the two already presented. The researchers used the transcripts of the interviews to classify subjects into epistemological understanding categories.

In order to make it feasible to work with younger subjects and larger samples, Kuhn and her colleagues developed a 15-item instrument designed to assess epistemological understanding across several domains (Kuhn et al., 2000). They report that a study using both the short form to assess epistemological understanding and the Livia task produced similar categorization of subjects (ibid). The short form will be used in this study and will be described in more detail in Chapter 4.

The link between epistemological understanding and reasoning has already been made in the domain of juror reasoning (Kuhn, 2001). The results of subjects who participated in both a juror reasoning task and an epistemological understanding task found that "the levels of [epistemological] reasoning on the Livia problem predicted performance on seven of the eight dimensions of juror reasoning" (Kuhn, 2001, pg. 6). Since this relationship may be the result of a third, lurking variable and not a causal relationship, correspondences between the two dimensions were investigated. Kuhn (2001) presents two examples of such correspondences, one from either end of the epistemological spectrum. One, an absolutist on the epistemological task, "demonstrated great trust in the absolute truth of a story in her reasoning as a juror" and "did not acknowledge the possibility of the evidence being used to tell another story (Kuhn, 2001, pg. 7)." Another, an Evaluativist in her epistemological task, "differentiated the evidence from theories of what might have happened and used the evidence critically to construct and evaluate theories in her juror's reasoning (ibid)."

The next chapter will present findings on human reasoning from the domain of psychology. These findings may be used to inform statistics educators of types of errors and misconceptions that may be developed or held by statistics students. In some of the most recent literature in the domain, psychologists study individual differences between subjects who differ in their responses to reasoning tasks. The findings of these studies will also suggest that Need for Cognition and Epistemological Understanding are suitable measures of productive dispositions in the learning of statistics.

Chapter 3: Literature Review

1. INTRODUCTION

Researchers in the field of psychology have been conducting studies designed to test the decision-making skills of humans since Peter Wason's work in the early 1960's (Evans and Newstead, 1995). In the early 1970's Tversky and Kahneman began to publish their now famous "heuristics and biases" literature. While researchers who have followed in the wake of Wason and Kahneman and Twersky have argued with the original findings (Hertwig and Gigerenzer, 1999), the fact remains that subjects are prone to making errors in judgment on the designed tasks (Stanovich, 1999). Psychologists continue to study these issues because they have come to believe that a modern technological society tends to promote human irrationality. Therefore, they conclude that an understanding of society in general is reliant upon an understanding of human reasoning (Stanovich, 1999).

It is the claim of this chapter that statistics educators should study these issues because the "human irrationality" exhibited by the psychology subjects predicts difficulties that students may have in learning statistics in general hypothesis testing in particular. These errors, and subsequent work in the field of human reasoning, may shed light on misconceptions statistics educators can expect from their students. Finally, recent research on psychological factors that tend to produce differences in responses to reasoning justify the choice of Need for Cognition and Epistemological Understanding as dispositions productive to the development of statistical proficiency.

In the decades since Wason and Kahneman and Tversky's original works, many studies designed to understand the process of human reasoning have been conducted.

The statistics education community can benefit from the results obtained by psychologists in studies of human rationality because the results will inform us in several areas. The errors made by psychology subjects may be used to predict errors and misconceptions that will be held by statistics students. Section two of this chapter describes some of the errors made by psychology subjects and explains how each of the errors discussed predict misconceptions and errors our students are likely to make. Research on the effects of changing the format of items can inform statistics instructors about versions of isomorphic questions that may be more difficult for students to answer correctly. Research pertaining to question format and on the use of simulations and the training of students to use representations to solve probability and statistics problems learning are presented in Section Three of this chapter. In addition, the possible effects of training activities on student learning are discussed. The chapter concludes with a discussion of the research on individual differences between psychology subjects. This discussion provides insight into the types of dispositions or world-views held by students that may promote success in learning to become statistically literate.

Stanovich (1999) defines normative models of reasoning to be those deemed to be logically or probabilistically correct. Throughout this chapter, the word "normative" is used to describe the response to a problem that is probabilistically, logically, and/or statistically correct. Non-normative reasoning is used to describe responses that are not the normative response.

2. SOME IDENTIFIED HEURISTICS AND BIASES

2.1 Errors Relating to Probability

2.1.1 Conjunction Fallacy

One of the irrationalities in reasoning enumerated by Kahneman and Tversky was the "conjunction fallacy" in which subjects rated a conjunction of the form, A and B, as more likely than one of the statements that comprised the conjunction when it stood alone (Tversky and Kahneman, 1983). Since the conjunction, A and B, is a subset of both set A and set B, it necessarily has a probability that is at most equal to the probability of the smaller of the two sets. Therefore, the probability of the conjunction statement must be less than or equal to the probability of each individual statement.

Kahneman and Tversky designed two problems to test for this particular irrationality, the Linda Problem and the Bill Problem. As the two are isomorphic, only the Linda Problem is discussed.

Linda Problem (Original)

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Rank the following statements by the degree to which Linda represents the typical member of that class:

1. Linda is a teacher in an elementary school.
2. Linda works at a bookstore and takes yoga classes.
3. Linda is active in the feminist movement.
4. Linda is a psychiatric social worker.

5. Linda is a member of the League of Women Voters.
6. Linda is a bank teller.
7. Linda is an insurance salesperson.
8. Linda is a bank teller and is active in the feminist movement.

Statement 8 is a conjunction of Statements 3 and 6 and, therefore, should be rated as less probable than either Statement 3 or Statement 6. However, approximately 80% of subjects, both statistically naïve and adept, rated Statement 8 as more likely than statement 6 (Kahneman and Tversky, 1982).

Misconceptions underlying the Conjunction Fallacy could cause statistics students difficulty in responding to probability questions that are considered to be fairly standard in textbooks for the introductory courses. Consider, for example, the following question from Moore's The Basic Practice of Statistics, a textbook designed for use in two and four-year universities and accessible to students with limited quantitative background, in other words, algebra- and not calculus-based (Moore, 2003).

> **Income and Savings**: A sample survey chooses a sample of households and measures their annual income and their savings. Some events of interests are
>
> A = the household chosen has income at least $100,000
>
> C = the household chosen has at least $50,000 in savings
>
> Based on the sample survey, we estimate that P(A) = 0.07 and P(C) = 0.2.
>
> a. We want to find the probability that a household has either income at least $100,000 *or* savings at least $50,000. Explain why we do not have enough information to find this probability. What additional information is needed?
>
> b. We want to find the probability that a household has either income at least $100,000 *and* savings at least $50,000. Explain why we do not have enough

information to find this probability. What additional information is needed? (Moore, 2003, pg. 300)

Students who exhibit the Conjunction Fallacy may not understand that these two sets may overlap and that the presence of elements in the intersection will effect the calculation of the probability of both the *and* and the *or* categories.

2.1.2 Conditional Probability

Casscells, Schoenberger and Grayboys created the following problem in 1978 for a study at the Harvard Medical School (Cosmides and Tooby, 1996). Their results, and the research that followed, can inform statistics educators of other types of errors students might be prone to making.

Medical Diagnosis Problem (Original form):

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about a person's symptoms or signs? ___%

Of the sixty medical students and faculty used as subjects in the original study, only 18% gave the normative answer of 2% (Cosmides and Tooby, 1996). Forty five percent of their subjects responded that 95% of those who test positive actually have the disease, appearing to have used the complement of the false positive rate as the true positive rate.

The following problem appears in Chapter 11 of The Basic Practice of Statistics (Moore, 2003, pgs. 295 - 296):

Testing for HIV

Enzyme immunoassay (EIA) tests are used to screen blood specimens for the presence of antibodies to HIV, the virus that causes AIDS. Antibodies indicate the presence of the virus. The test is quite accurate but not always correct. Here are

the approximate probabilities of positive and negative EIA outcomes when the blood tested does and does not actually contain antibodies to HIV.

	Test Results	
	+	-
Antibodies Present	0.9985	0.0015
Antibodies Absent	0.0060	0.9940

Suppose that 1% of a large population carries antibodies to HIV in their blood.

Use the information in the previous exercise and the definition of conditional probability to find the probability that a person has the antibody, given that the EIA test is positive.

Given the fact that medical students and faculty fare so poorly with the same question, even though the question is in a domain with which doctors should be familiar, it seems unreasonable to expect undergraduate students to answer this type of question in a beginning statistics course. Later in this chapter, further work using the Medical Test Problem will be reviewed and the findings will be used to suggest a more appropriate phrasing of this question for the population of students in an introductory statistics class.

2.2 An Error in Testing Hypotheses – Pseudo-diagnosticity

Another example of non-normative reasoning found in the psychology literature that may affect a student's ability to master statistical proficiency, particularly in the domain of hypothesis testing, is called "pseudo-diagnosticity" (PD). When making this error in judgment, subjects who are given a choice among various pieces of evidence that will help in a decision-making task tend to request one that is not optimal from a diagnostic standpoint. The level of diagnosticity of evidence that is presented to a decision maker who is choosing among alternatives in a judgment task may be quantified

by the degree to which the evidence discriminates between the available hypotheses (Evans, Venn and Feeney, 2002). Information about only one of the hypotheses is said to be "pseudo-diagnostic" because, while it may provide evidence for or against that hypothesis, it does not help discriminate among the hypotheses. Empirical findings that have been replicated in many studies suggest that the modal response given by subjects on PD tasks is a pseudo-diagnostic request for information.

A standard item for testing the propensity of subjects to exhibit pseudo-diagnosticity (PD), taken from Mynatt et al. (1993), is reported by Evans, Venn and Feeney (2002):

> **Sister's Car Problem (Original):**
>
> Your sister has a car she bought a couple of years ago. It's either a car X or a car Y but you can't remember which. You do remember that her car does over 25 miles per gallon (mpg) and has not had any major mechanical problems in the two years she has owned it.
>
> You have the following information:
>
> A. 65% of car X's do over 25 miles per gallon
>
> Three additional pieces of information are also available:
>
> B. The percentage of car Y's that do over 25 miles per gallon
>
> C. The percentage of car X's that have no major mechanical problems for the first two years of ownership.
>
> D. The percentage of car Y's that have no major mechanical problems for the first two years of ownership.

The piece of information that is diagnostic in the sense that it provides information needed to discriminate between car X and car Y is B: the percentage of car Y's that do over 25 miles to the gallon. If we know that piece of information we can compare cars X

with Y over the factor of gas consumption. The car that tends to have better gas consumption is more likely to be the car owned by the sister. From a Bayesian perspective, we would have enough information to create a complete likelihood ratio required to make a probabilistic statement about the car type (Evans, Venn and Feeney, 2002).

The modal choice by subjects is to request item C: the percentage of car X's that have no major mechanical problems for the first two years of ownership. This option is "pseudo-diagnostic" because it, together with the first piece of information, provides the subject only data about car X's. The subject has no way to discriminate between X and Y because she has no basis for comparison; items A and C cannot be used to create a likelihood ratio involving both types of cars. It is hypothesized that subjects choose this information because the first piece of information focuses the attention on car X, and they are trying to confirm a hypothesis about car X without regard to an alternate hypothesis about car Y.

In order to further investigate these ideas, Evans, Venn and Feeney (2002) devised a new task with three versions, manipulating the first piece of information given to the subjects.

Sister's Car Problem (Initial Information Manipulation)

> Your sister bought a new car. You can't remember whether it's a model X or a model Y, but you do remember that two of the things she is concerned about are petrol consumption and good mechanical reliability. We have already asked the following question for you and have given you the answer.
>
> Question A
>
> What is the average miles per gallon achieved by model X cars?
>
> Answer: (three versions) 20 mpg/35 mpg/50 mpg

You may learn the answer to one more question. Circle your choice.

B. Average mpg of car Y's

C. Percentage of X's with good mechanical reliability.

D. Percentage of Y's with good mechanical reliability.

The participants who were initially given the information that car X's average 20 mpg were less likely than the other subgroups to request pseudo-diagnostic information (option C) as their second question. They were not more likely, however, to ask for the normative diagnostic information than were the subjects assigned to other conditions. Rather, they were more likely to ask for the information provided by option D: the percentage of car Y's with good mechanical reliability. It is hypothesized that the fact that car X's have poor gas consumption convinced these participants that the car was a Y (or perhaps that it was not an X). Not only was the attention of the subjects refocused on the hypothesis that the car is a Y, but the subjects appeared to assume that since car X achieves poor gas mileage, that car Y achieves good gas mileage. This led the participants to request information regarding car Y on the dimension about which the subjects had not already made assumptions (Evans, Venn and Feeney, 2002).

These findings are troubling in terms of the ability of students to master the concept of hypothesis testing for several reasons. First, they indicate that students may enter a statistics course having poor notions of what types of evidence they might choose to evaluate the truth of a statement or to discriminate between two possible statements. This suggests that students are ill prepared to learn statistical reasoning, lacking the ability to evaluate standards of evidence. Furthermore, according to Evans, Venn and Feeney (2002), "there is convincing evidence in the literature that people habitually think about only one hypothesis at a time" (pg. 40). This propensity may pose a problem for

beginning learners of hypothesis testing. The first step in performing a formal hypothesis test in an introductory statistics class is to formulate a set of two hypotheses, the null hypothesis and the alternate hypothesis. The difficulty presented by the naming of two hypotheses is further complicated by the fact that the first hypothesis to be written, the null hypothesis, is the opposite of the hypothesis for which the student is asked to find evidence. It is likely that this process, which forces the student to consider two hypotheses simultaneously in a manner that is not intuitive, may create a large cognitive load for that student. This feature of the process would, therefore, add to the overall difficulty of mastering the formal process.

A typical task in an introductory statistics class designed to test whether a student can carry out a hypothesis test might be: "In a simple random sample of 750 likely voters, 380 said they would vote for Kerry, 355 said they would vote for Bush and 15 said they were still undecided, if the election were held today. Do the data presented give good evidence that Kerry has a majority of the popular support at this time? Carry out a test and give conclusions." The first step in the correct response to the question of whether Kerry has a clear majority is to write down, using mathematical symbols that he does NOT have a clear majority. The next step is to write in mathematical symbols that Kerry does have a clear majority. While "the evidence falls well short of convincing us that people cannot consider alternative" hypotheses (Evans, Venn and Feeney, 2002, pg. 40), it seems reasonable to believe that students would need specific training in order to complete the formal process of writing hypotheses without losing focus as a result of the cognitive overload that may be induced by the first step of the process.

2.3 Errors relating to difficulties that result from subjects' lack of expertise in navigating between contextual information and statistical information

Recall from the previous chapter that strategic competence in statistics includes the ability to navigate correctly between statistical information and contextual information. The errors discussed in this section all stem from subjects' difficulty with this aspect of proficiency. The biases described can all be traced to subjects' over reliance on contextual information and a lack of use of statistical information. Two such biases, belief bias and framing effects bias, are described in this section. The general class of biases that result from inexpert incorporation of contextual data is called Fundamental Computation Bias (FCB). The general description of FCB is beyond the scope of this work.

2.3.1 Belief Bias

Belief Bias, the predisposition of subjects to be influenced by beliefs that are not logically relevant to the given task (Evans, Brooks, and Pollard, 1985), arises in argument validation studies. Evans, Barston and Pollard (1983) asked subjects to decide whether a given conclusion could be logically deduced from the given premises. There were four argument types: 1) valid with believable conclusion, 2) valid with unbelievable conclusion, 3) invalid with believable conclusion, and 4) invalid with unbelievable conclusion. The study found main effects for validity and believability – subjects were more likely to endorse a conclusion when the reasoning was valid rather than invalid and when the conclusion was believable rather than unbelievable. This latter effect exemplifies Belief Bias. There was also a significant interaction between the two variables in that subjects were much less likely to endorse the unbelievable conclusion in

the invalid reasoning case than in the valid reasoning case. In addition, Thompson (1996) found that believability of premises affects subjects' rating of strength of argument.

In summary, the literature on Belief Bias in reasoning indicates that subjects' ratings of strength of arguments are affected both by the believability of conclusions and the believability of the premises. Beginning statistics students learning hypothesis testing are essentially learning a new type of argument from which to make conclusions. Before this process is accepted as valid by students, it is possible that conclusions that are counter to the students' beliefs would encourage students to suspect that the process that produced them is invalid. If this is the case, using teaching examples that produce believable or confirmable results should increase student belief in the validity of the process while using examples that lead to unbelievable conclusions should decrease student belief in its validity. Furthermore, the Belief Bias studies predict that students who do not believe in the validity of the process of hypothesis testing, in other words, those who have not developed a certain level of conceptual understanding of the topic, may tend, in large numbers, to reject unbelievable conclusions produced by the process.

2.3.2 Framing Effects

First proposed by Kahneman and Tversky in a 1981 publication (Kuhberger, 1998), the Framing Effects Bias predicts decision-making behavior under particular constraints. More specifically, when a subject is presented with two choices, one with a certain outcome and the other with a "risky" or probabilistic outcome, she will tend to make a different decision based on the framing of the outcomes. According to Kuhberger (1998), Kahneman and Tversky employed a strict definition of framing, that is, framing is a semantic manipulation of outcomes that are formally identical. In their seminal paper on framing effects, Kahneman and Tversky used two outcome sets, each with two

choices. In one outcome set, the positive aspect of the outcomes was made salient; in the other the negative aspect of the outcomes was highlighted. All four choices had identical expected values from the standpoint of probability theory.

The now famous example that initiated research on framing effects illustrates the theoretical construct:

Asian Disease Problem:

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows. Which option would you choose?

The positively framed set of outcomes:

- If Treatment A is utilized, 200 out of 600 people are certain to recover from the outbreak of disease X.
- If Treatment B is utilized, there is a 1/3 chance that all 600" people will recover and a 2/3 chance that no one will recover.

The negatively framed set of outcomes:

- If Treatment A is utilized, it is certain that 400 out of 600 people will die from the disease.
- If Treatment B is utilized, there is a 1/3 chance that no one will die and a 2/3 chance that everyone will die from the disease.

Subjects were presented with one set of options from which to choose. In each set of options listed, the certain outcome is the first choice and the probabilistic outcome is the second choice. In each of the four options, the expected value of lives saved is 200 and the expected values of lives lost is 400, so there was no reason to believe that subjects would respond differently to the two tasks. Kahneman and Tversky found, however, that while 72% of the subjects who were presented with the positively framed

outcomes chose the certain outcome, 78% of the subjects who were presented with the negatively framed outcomes chose the probabilistic outcome (Miller and Fagley, 1991).

In 1998, Kuhberger completed a meta-analysis of the 152 empirical papers that had been published on Framing Effects Bias. He found that some of the experiments used what he calls a "loose" definition of framing. Under the loose definition of framing, framing "can be induced not only by semantic manipulations but may result also from…contextual features of a situation and individual factors, provided that the problems are equivalent from economic theory" (Kuhberger, 1998, pp. 24). The results of the meta-analysis corroborate the existence of framing effects with a moderate effect size. Kuhberger concludes, however, that there are certain semantic and task manipulation factors that mitigate framing effects (Kuhberger, 1998). Notable is the fact that the Asian Disease problem one of the better tasks at producing Framing Effects Bias.

The existence of a bias such as framing effect highlights the difficulty people have in properly integrating contextual information and statistical information. This particular bias serves as a reminder the importance of context on decision making even when the statistical outcomes presented are equivalent. The ramifications of the importance of context in the development of strategic competence for beginning students of statistics should guide the selection of examples and exercises used in the classroom and as assessment. The consequences of the importance of context and the difficulties students have in correctly incorporating contextual and statistical information will be discussed in more detail later in this work.

3. MANIPULATIONS THAT PROMOTE NORMATIVE REASONING

This section describes three types of intervention studies in which manipulations improved subjects' correct response rate. The types of studies are, 1) those in which the

question was written in a different format, 2) those that required the subjects to perform simulations designed to confront known misconceptions, and 3) those that provided subjects with a representation designed to help think about the problem normatively.

3.1 Changing the Question Format – Frequency Format and Distributional Form

This section discusses how changing the format of some of the questions described above results in a significantly higher percentage of subjects giving the correct response. The manipulations provide insights for statistics educators writing assessment items for their student.

3.1.1 The Linda Problem

One might conclude that the conjunction fallacy exhibited by subjects in the Linda Problem is the result of the subject is allowing the context of the problem to override the logical structure. That is, they are basing their judgments on the context of the description rather than considering the mathematical structure of the problem. Work that follows that of Kahneman and Tversky, however, shows otherwise.

One of the theories about the mechanism underlying the conjunction fallacy was that subjects were having difficulty in applying probability in the singular case (Cosmides and Tooby, 1996). From the definition of probability given in a standard introductory statistics text, which is that probability is the proportion of times the outcome would occur in a very long series of repetitions (Moore, 2003), it would not be surprising if the consideration of the expected value from a sample leads to more normative responses than does the probability of a single outcome. From a frequentist view, Linda either is or is not a bank teller and Linda either is or is not a feminist. We cannot predict this single

outcome, we can only report the propensity of the possibilities associated with the sample space of outcomes.

> In fact, when Fiedler modified the Linda task so that the question statement read,
>
> To how many out of 100 people who are like Linda do the following statements apply?

he found that 70 – 80% of subjects did not make the conjunction fallacy (Cosmides and Tooby, 1996). This suggests that subjects will reason more normatively when cued to think about a single event as one part of a larger possible sample space.

Returning to the Income and Savings problem discussed in the original section about the Conjunction Fallacy, the results of Cosmides and Tooby (1996) suggest that instruction of probability of *and* and *or* cases may benefit from specific use of enumerated cases. For example, after the probabilities associated with income and savings are given, students might be asked how many families out of 100 would they expect in each category. Then the question might be posed about the total number of families in both categories. Cosmides and Tooby's research suggests that, in the process of enumerating the actual families, students may be prompted to realize that some of those families overlap.

3.1.2 The Medical Test Problem

Further research on the interaction between question format and probabilistic reasoning found that posing the question using a frequency format rather than as a percentage improves normative performance. Cosmides and Tooby (1996) published a major work in this branch of the literature using variations of the Medical Test Problem stated above. Cosmides and Tooby (1996) first replicated the findings of Casscells et al. using Stanford undergraduates as subjects. Subsequently, they found that they could

significantly raise the rate at which respondents gave the normative answer by changing the presentation of the problem to one that gave information and requested a response in frequency format rather than via a percentage:

> **Medical Diagnosis Problem (Frequency form with explanation):**
>
> One out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease. Imagine that we have assembled a random sample of 1000 Americans. They were selected by a lottery. Those who conducted the lottery had no information about the health status of any of those people.
>
> Given the information above:
>
> On average,
>
> How many people who test positive for the disease will *actually* have the disease? ___ out of ____?

This version of the problem differs from the original in more than that a frequency presentation is used and that an answer is requested in frequency format. This version also explains the concept of a false positive and specifies the true positive rate. Furthermore, it highlights the salience of the random sample and makes the sample more concrete by enumerating its size (Cosmides and Tooby, 1996). Fiedler demonstrated that the enumeration of cases would improve the normative response rate remarkably, by itself. In order to unpack the possible reasons for the rise in correct response rate, Cosmides and Tooby followed up with a study that used a similar text to the frequency form with explanation except that the data were presented in percentage format together with the same explanation that was given in the frequency format with explanation version.

Medical Diagnosis Problem (Probability format with explanation)

> The prevalence of disease X is 1/1000. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, 5% of all people who are perfectly healthy test positive for the disease.
>
> What is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the symptoms or signs?_____%

Subjects did significantly better (36% gave a normative response) on this version of the question than those who responded to the original version. This indicates that the explanation of false positive does aid performance. The percent of correct responses to this version, however, was significantly lower than on the frequency format with explanation demonstrating that the frequency format encourages normative performance.

One might note that it is possible that Cosmides and Tooby have merely replicated Fiedler's findings that a subject is more likely to give the normative response when the sample is salient. In the probability format version of the above problem, the question asks about the probability that 1 man who tests positive actually has the disease, whereas in the frequency format version the question posed asks about the number of people in a large group who one would expect to have the disease. In order to test the effect of the specificity of the sample, the following version, in which a probabilistic sample was specified, was also tested:

Medical Diagnosis Problem (Probability format with explanation and random sampling assumption):

> The prevalence of disease X among Americans is 1/1000. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test

> also comes out positive when it is given to a person who is completely healthy. Specifically, 5% of all people who are perfectly healthy test positive for the disease.
>
> Imagine that we have given this test to a random sample of Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.
>
> What is the chance that a person found to have a positive result actually has the disease? _____%

Twenty-eight percent of the subjects gave the correct Bayesian response. Sixteen percent gave the base rate response of 1/1000. These results are not significantly different from the results found using the **"**probability format with explanation" version of the problem. The difference between the two versions is the salience of random sampling. Since there were no significant differences in performance on the two items, it appears that the salience of random sampling is not a factor in subjects' ability to respond to this type of item.

Girotto and Gonzalez (2001) used an item similar to the Cosmides and Tooby item in order to further assess the effect of reasoning about a singular case. Each subject solved two problems. One was presented in frequency format and the other in chance format. Girotto and Gonzalez realized, however, that in each of the preceding examples, the frequency format requests a two step response, providing a space for both the numerator and the denominator of the ratio, while the probability format requests a one-step response. Therefore, they developed both a one-step and a two-step response question for each format. Each participant was asked the same type of question for each of the two formats (one of: two-step chance, two-step frequency, one-step chance and one-step frequency).

Frequency version:

4 out of 100 people tested were infected.

3 out of 4 infected people have a positive reaction to the test.

12 of the 96 uninfected people also had a positive reaction to the test.

Chance version:

A person who was tested has 4 chances out of 100 of having the infection.

3 out of 4 chances of having the infection were associated with a positive reaction to the test.

12 of the remaining 96 chances of not having the infection were also associated with a positive reaction to the test.

Two step Chance version:

Imagine that Pierre is tested now. Out of a total of 100 chances, Pierre has _____ chances of having a positive reaction, _____ of which will be associated with having the infection.

Two step frequency question:

Imagine that a group of people is tested. In a group of 100 people, one can expect _____ individuals to have a positive reaction, _____ of whom will have the infection.

Single-step chance question:

If Pierre has a positive reaction, there will be _____ chance(s) out of - _____that the infection is associated with his positive reaction.

Single-step frequency question:

Among 100 people who have a positive reaction to the test, the proportion that has the infection will be equal to _____ out of _____.

Subjects correctly solved two-step questions at a higher rate than the one-step versions, regardless of format of presentation or response. This indicates that drawing attention to both parts of the ratio increases the ability of subjects to give the normative response.

The research results discussed in this section suggest that rewriting the Testing for HIV question, cited above from an introduction to statistics textbook, as stated below might help students obtain the correct response at a higher rate. While the development of procedural fluency only as a result of certain prompts is not the ultimate goal of a statistics course, the use of such examples and assessments when first introducing topics may aid students in developing confidence in their abilities to master the topics.

Testing for HIV

Enzyme immunoassay (EIA) tests are used to screen blood specimens for the presence of antibodies to HIV, the virus that causes AIDS. Antibodies indicate the presence of the virus. The test is quite accurate but not always correct. Here are the approximate probabilities of positive and negative EIA outcomes when the blood tested does and does not actually contain antibodies to HIV.

Out of every 10,000 people who have the antibodies, 9985 will test positive for the antibodies and 15 will test negative. Out of every 10,000 people who do not have the antibodies, 60 will test positive and 9940 will test negative. 10,000 out of every 1,000,000 people carry the antibodies to HIV.

Imagine that a group of people is tested. In a group of 1,000,000 people, one can expect _____ individuals to have a positive reaction, _____ of whom will have the infection.

Use your response above to find the probability that a person has the antibody, given that the EIA test is positive.

The conclusions from this branch of research on human reasoning, that subjects reason more normatively with data given in frequency format and when the instance as

part of a larger possible sample is made salient, may also have direct consequences on the learning of hypothesis testing in statistics. The first claim suggests that, when introducing hypothesis testing about proportions, data should be given in frequency format, as counts, rather than as percentages. Furthermore, while it may not appear to follow directly from the theory posed above, it is likely that students can reason more normatively about significance testing regarding proportions than they can about those regarding means. Indeed, this claim will be supported in a later discussion of successful training techniques that promote normative reasoning. It will be shown that students can create more useful representations, strengthening strategic competence in statistics, when solving problems with data given in frequency format.

The second finding of the cited research on question formats, that the salience of an individual case as part of a larger possible sample promotes normative reasoning, has implications for the growth of conceptual understanding of hypothesis testing. A p-value is the probability of obtaining a test statistic as unusual or more unusual than the one we derived from our sample if the null hypothesis was true. We have taken one sample, but we are reasoning probabilistically about that sample as a representative of the larger population of all samples of that size. If students do not view the sample chosen as one of many possible samples, they may not understand the difference between the results of a random sample and an anecdote. Each will be viewed as a singular, rather than distributional, event. In order to make salient to beginning students the idea that a sample is one of many possible samples, perhaps we should initially phrase questions as, "Use the p-value to express the chances that a sample chosen at random would have a mean as high or higher than the one our sample had if the null hypothesis were true. _____ out of _____ samples" or "If we chose 100(0) random samples, how many of them would we

expect to have a mean this high if the null hypothesis were true?" This would mirror the model provided by Girotto and Gonzalez and allow the salience of the probability of sampling to become apparent.

3.2 Simulations Training

In this section we discuss two examples of simulation training in which subjects use simulations to face misconceptions in reasoning. Researchers in psychology conducted the first study using undergraduate subjects. Statistics educators conducted the second using statistics students. The students worked with the simulations as part of their coursework in a statistics class.

3.2.1 The Monty Hall Problem

The Monty Hall Problem (MHP) is problem from the psychology literature that has associated with it a tradition of remediation via simulations designed to confront the commonly held misconceptions. The original version of the problem was adapted from the TV game show Let's Make a Deal. In the final challenge each week, a contestant was shown three doors: one held a prize and the others held nothing (or perhaps a goat). After the contestant chose a door, the host of the show, Monty Hall, would reveal one of the two remaining doors. At this point, the contestant was given the choice to switch from her original door choice to the other remaining door.

If we assume that Monty Hall has chosen an empty door to open, then probabilistically, the correct answer, if one is interested in maximizing one's chance of winning, is to switch doors (Bar-Hillel and Falk, 1982). When there were three doors available, each door held a 1/3 chance of being a winning door. When a losing door is revealed, the originally chosen door still holds a 1/3 chance of winning and the remaining door, the one not originally chosen, now has a 2/3 chance of being the winner. The

modal response by psychology subjects is to assert that each door has an equal probability of winning, (Franco-Watkins, Derks, and Dougherty, 2003; Tubau and Alonso, 2003), so that the decision to stay or switch is a matter of personal preference and makes no difference in the long run. To gain an intuition of why the decision to switch is the normative response, imagine that the game had 100 doors and the contestant chose one. The chosen door has a 1/100 probability of containing the prize. When he opens the ninety-eight remaining doors that do not hold prize, the originally chosen door still has a 1/100 chance of being the door with a prize.

Both Franco-Watkins, et al. (2003) and Tubau and Alonso (2003) reported the results of training activities on the choices subjects make in subsequent playing of the Monty Hall game. In both studies, subjects played repeated trials of the Monty Hall game on the computer, some of the subjects were trained using a 10 choice version and some using the standard 3 choice version. They were told that the goal was to find the prize and were given the opportunity to stay with or switch from their original choice after all the remaining doors were opened. In both versions, the subject had two doors left from which to choose, his original door and one other. In the 3-door version, one door was opened; in the 10-door version, 8 doors to empty rooms were opened. The computer recorded the results for switching versus staying so that the subjects could learn from the pattern of playing.

The findings of both research groups were roughly equivalent; through repeated trial in the simulation, subjects learned to switch doors. Franco-Watkins et al. tracked switching behavior over sets of trials and the percent of subjects who gave the normative response rose over time whether the simulation was based on a three-door game or a ten-door game. While the subjects learned the normative procedure, they did not develop

understanding about the situation (Franco-Watkins, Derks, and Dougherty, 2003; Tubau and Alonso, 2003) as they "overwhelmingly judged the probabilities as .50 regardless of whether they stayed or switched" (Franco-Watkins, Derks, and Dougherty, 2003, pg. 77). These particular training studies, therefore, show an improvement in procedural fluency without a corresponding improvement in conceptual understanding.

3.2.2 Sampling Distributions

The Monty Hall problem might be considered to be a contrived problem with little applicable value. The results, however, that were found in the Monty Hall simulation study described above were similar to those found by an instructional intervention that used simulations to help statistics students learn about sampling distributions. The instructional intervention was reported by delMas, Garfield and Chance (1999). The results of the instructional intervention with regard to conclusions about the development of procedural fluency and conceptual understanding were similar to those found by the Monty Hall simulation study. The students appeared to develop procedural fluency without developing conceptual understanding.

In order to help students develop deeper understandings of the concepts of sampling distributions and the central limit theorem, delMas, Garfield and Chance (1999) developed a computer program and instructional materials that were used by beginning statistics students as part of their first course in statistics. The protocol was a pre-test-post-test design and encompassed three cohorts of student subjects. Between the cohorts, the simulation program was redesigned to address more fully the misconceptions that remained in the post-test results of the previous cohort. In general, the program allowed students to generate many random samples from various shaped distributions and record the parameters of the distribution as well as the statistics from the samples. In the second

and third cohort activities, students were asked to make predictions before selecting samples in order to address misconceptions explicitly.

While the scores on the post-tests were higher than the scores on the pre-test for all cohorts involved in the delMas, Garfield and Chance (1999) research, the authors were "disappointed to find that many...students still displayed some serious misunderstandings of sampling distributions" (pg. 11) even after having completed the simulation activity. As with the MHP simulation, it appears that, while the sampling distributions simulation improved students' procedural fluency, it did not address issues required for growth of conceptual understanding.

3.3 Representations Training

Another form of training study is based on the cognitive perspective that claims when subjects can create a representation of a situation, they are more likely to compute the correct response to the item. The "mental models" theory of human reasoning specifies that the process of reasoning includes, constructing a model (or set of models) based on the premises and using general knowledge to make explicit something only implicit in the premises (Johnson-Laird, 1994). Thus, the ability to create a representation of the problem appears to be a crucial element leading to a correct solution. This section will discuss two studies in which subjects were taught new representations to aid in solving previously difficult problems.

3.3.1 The Medical Test Problem

In order to create an item that could be represented more easily, Cosmides and Tooby revised their medical diagnosis item, discussed above, so that

- the sample size was 100 (disease rate was 1/100 and false positive rate was 5/100)

- subjects were given a 10 x 10 grid representing a sample of 100 people
- subjects were required to circle the boxes representing people who had the disease and to fill in the number of people who would test positive *before* they gave the rate of positive tests that actually indicate the disease.

Under these conditions, 92% of subjects gave the normative response to the item, significantly higher than on any other administration of the task (Cosmides and Tooby, 1996). In fact, the authors also found that many of the subjects who gave the normative response to the percent version of the problem with explanation of false positive left evidence in their booklet of having enumerated the cases of positive tests in a manner congruent to the representational prompts provided in the study cited above. While this item does not reveal anything about the subjects' conceptual understanding or the development of their reasoning skills, it does indicate that the ability to create representations, a strategic competence, enhances procedural fluency, the ability to provide the normative response.

The HIV Test question from an introduction to statistics textbook cited above follows a section devoted to creating tree diagrams and, in one of its parts, directs the student to create a tree diagram (Moore, 2003). The findings of Cosmides and Tooby (1996) on the use of the grid model of representations suggests that, perhaps, the inclusion of that model in addition to the tree diagram might aid in statistics students' development of procedural fluency in conditional probability.

3.3.2 Probability and Hypothesis Testing – Grids, Venn Diagrams and the Flexible Urn

The notion that training students to create suitable representations in order to complete a standard probability task was also the subject of a research project conducted

by Peter Sedlmeier. Sedlmeier (1999) provided training for students using either a Venn diagram or a grid approach for probability questions and a flexible urn model for questions about sampling distributions. In the flexible urn model, subjects were trained to imagine the sampling step of a hypothesis test process as if they were taking a random sample of balls from an urn. In the case of a test for proportions, the balls were thought to be of two colors. In the case of a test for means, the balls were considered to have values.

Sedlmeier (1999) found, for all three representations, that the training was successful on transfer problems and in follow up tests five weeks after the training. Further, the representations training was significantly more successful than rule-based training. As with the Cosmides and Tooby representations training study discussed above, a shortcoming of the Sedlmeier work is that the post tests measured only procedural knowledge so there is no way to know whether the representations training is effective in helping students develop stronger conceptual understanding.

4. STUDIES OF INDIVIDUAL DIFFERENCES

The psychology literature on human reasoning discussed to this point has two main goals. First, using the inconsistencies exhibited by subjects as a basis, researchers attempted to create a model for the reasoning processes of humans. One outcome of this body of literature is Sloman's dual processor model of reasoning (Sloman, 1996). The second goal, as we saw in the previous section, was to create training and remediation interventions to teach subjects to reason normatively. Stanovich and West, however, looked at this domain through a different lens. These researchers noticed that there are subjects who reason normatively, even on the most abstract and difficult formulations of the clinical tasks (Stanovich and West, 1998b). They set out to identify factors, or traits,

that predict success in the difficult tasks. These factors are called "individual differences" in the literature. The results of studies and that attempt to find individual differences that affect the process of human reasoning are discussed in this section.

4.1 General Reasoning Tasks

The large-scale research project completed by Stanovich and West comprised more than 1500 subjects and used items that tested many of the aforementioned biases: belief bias, pseudo-diagnosticity, statistical reasoning, framing effects, and conjunction bias (Stanovich and West, 1998a, 1998b, 1997). Briefly, the premise of the research is that differences in performance on the human reasoning tasks will be accounted for by two separate yet related factors, cognitive capacities and thinking dispositions, and that each factor will affect a different aspect of human reasoning.

Cognitive capacities are defined within this research program as those factors that "underlie traditional psychometric intelligence" (Stanovich and West, 1998b, pg. 163), such as perceptual speed and working memory capacity, which are thought to be improved by long term practice rather than short term instruction. In contrast, thinking dispositions are "related to the adequacy of belief formation and decision making" (ibid) and tend to be more malleable in nature than cognitive capacities. The authors hypothesize that differences in cognitive capacities lead to differences in algorithmic processing, while differences in thinking dispositions lead to differences in adaptive reasoning skills.

In order to quantify cognitive capacities, Stanovich and West used a composite measure based on "well-known cognitive ability and academic aptitude tasks" (1998b, pg 165), such as self-reported SAT scores and reading comprehension tasks. Thinking

dispositions were captured through a composite measure accounting for open-minded and counterfactual thinking as well as absolutism, dogmatism, paranormal beliefs and social desirability. Through their work, Stanovich and West found significant correlations between cognitive capacities and thinking dispositions. Further, they found that the capacities and dispositions were correlated with differences in performance on the typical psychological reasoning tasks. In addition, each composite measure, capacity and disposition, was found to be a significant, unique predictor of differences in normative reasoning, supporting the authors' theory that differences are affected by at least two distinct factors: capacity and disposition.

Stanovich and West's thinking dispositions composite score (TDC) was a linear combination of subjects' scores on five subscales. The scores on the actively open-minded thinking and counterfactual thinking scales had positive factor loadings and scores on Absolutism, Dogmatism and Paranormal scales had negative factor loadings (Stanovich and West, 1998b). The authors claim that high scores on the TDC "indicate open-mindedness, cognitive flexibility, and a skeptical attitude" (Stanovich and West, 1998b). In considering the Actively Open-Minded Thinking and Absolutism scales, it appeared as though those tasks measured ways of thinking and viewing knowledge rather than a personality trait. Furthermore, while Stanovich and West have enumerated some factors that explain individual differences in human reasoning, they have not accounted for all the variability that is found in subjects. Epistemological Understanding, a construct discussed in Chapter Two, presents another way of measuring ideas about knowledge, so it makes a reasonable substitution for two of the five subscales used by Stanovich and West. Need for Cognition, as discussed in Chapter Two, has also been associated with personality types that describe successful statisticians, and its questions more specifically

address the dispositions described than do the Stanovich and West subscales. In the next two sections we will also see that two constructs, Need for Cognition and Epistemological Understanding, have been associated with differences in reasoning.

4.2 Framing Effects Bias and Need for Cognition

A general argument for the use of Need for Cognition (NC) as a measure of productive disposition in statistics learning was made in the previous chapter. It is also the case that the relationship between NC and individual differences in human reasoning has been investigated directly in the psychology domain. At this time, it appears that NC interacts with framing effects in decision making. Recall that a framing effect is said to occur when a subject is induced to choose a different outcome when the wording or context of the outcomes of a decision-making task is changed from highlighting positive outcomes to highlighting negative outcomes without changing the substance of the outcomes.

LeBoeuf and Shafir (2003) have shown that those high in NC are less likely to give inconsistent responses to two equivalent tasks, one with positive frame and the other with negative frame. That is, subject high in NC are more likely to select the risky or sure options in each of two tasks, one positively framed and one negatively framed, rather than switching between the risky and sure options when the framing is changed. This result indicates that NC does interact with the capacity for reasoning. It should be noted, however, that, while high NC subjects tend to be consistent across two tasks, they still exhibit framing bias based on which frame, positive or negative, is presented first.

Based on these initial findings, it seems reasonable to conclude that NC may be a dispositional factor that affects learning of statistical reasoning. Just as Stanovich (1999)

reports significant correlations between cognitive abilities and thinking dispositions, NC has been found to have a positive correlation with intelligence (Cacioppo and Petty, 1982) and achievement in school (Diseth and Martinsen, 2003). Therefore, any investigation of a relationship between thinking dispositions should take into account general academic achievement and intelligence in order to determine the contribution of these along with NC so that factors are not masked by correlates.

4.3 Juror Reasoning and Epistemological Understanding

Deanna Kuhn suggests that epistemological beliefs may help to augment the explanation provided by Stanovich of the individual differences encountered in human reasoning (Kuhn, 2001). Kuhn asserts that of the dispositional cognitive variables considered by Stanovich, epistemological understanding is the "most fundamentally clear" of the constructs (Kuhn, 2001, pg. 5). The description of the construct and the results of Kuhn's work on juror reasoning presented in Chapter Two of this work provide a further argument for its inclusion in the study of individual differences in student development of statistical proficiency.

This chapter began with the presentation of tasks, used by psychologists who study reasoning, that predict difficulties for beginning statistics students. Psychology researchers developed three main research domains from the original tasks: 1) they used manipulations of the tasks to develop a model to describe human reasoning; 2) they developed training interventions to produce more normative reasoning and; 3) they investigated individual differences that contribute to normative reasoning. It is a claim of this chapter that all three avenues of research can inform research and practice in statistics education. The study presented with this work, however, seeks to explore the implications of the individual differences research on statistics learning. In particular, the

study is designed to test the relationship between Need for Cognition and Epistemological Understanding and the development of statistical proficiency of beginning students of statistics.

Chapter 4: Experimental Design

1. OVERVIEW

The goal of this research is to develop a dispositional attribution model to explain individual differences in the development of statistical proficiency as described in previous chapters. Dispositional attribution is the act of explaining a behavior as the result of underlying psychological factors (Corsini, 1999). In this research, I attempt to explain differences in statistical proficiency using two psychological factors: Epistemological Understanding (EU) and Need for Cognition (NC).

The research began with a Pilot Study designed to test assessments of the psychological constructs and aspects of statistical proficiency as defined in Chapter 2. I sought to show that the responses exhibit enough variability to be worthwhile in finding differences between subjects. The next phase of the research, the Population Study, was designed to find a baseline distribution of the psychological factors in the target population, undergraduate students taking an algebra-based introduction to statistics course. The results of this study, conducted with a large sample, would be used to assess the independence of the two psychological constructs. The third phase of the project, called the Exam Study, allowed me to test several items that assess conceptual understanding of hypothesis testing on a large sample. In addition, this study allowed me to assess the relationship between the psychological constructs and the development of procedural fluency and conceptual understanding. Finally, this study provided information on the level of procedural fluency and conceptual understanding of subjects who would be chosen for the next phase of the study, the Targeted Study. In the Targeted Study, subjects from various categorizations of Need for Cognition and

Epistemological Understanding would be interviewed using items designed to assess strategic competence and statistical reasoning about hypothesis testing. Differences between groups of subjects would lay the foundation for the development of a model using the psychological constructs to explain differences in the development of statistical proficiency.

1.1 Assessing Epistemological Understanding

Epistemological Understanding (EU) was assessed using the 12-item instrument described in Kuhn, Cheney and Weinstock (2000). It was coded in the same manner that was specified in that paper. The 12-item instrument was designed to measure EU in four domains: aesthetic judgments, value judgments, judgments about truth in the social world, and judgments about truth in the physical world. It has three items pertaining to each domain. The items were not grouped by domain and the same order was retained throughout the set of studies described here.

Each item consists of a pair of competing assertions. The 12 items that comprise the assessment are reproduced in Appendix A. One example of an item designed to measure judgments about truth in the social world is,

- Robin thinks the government should limit the number of children families are allowed to have to keep the population from getting too big.
- Chris thinks families should have as many children as they choose.

For each set of assertions, subjects were asked to respond to the following questions:

A. Can only one of their views be right, or could both have some rightness? (CIRCLE ONE)

   i. ONLY ONE RIGHT ii. BOTH COULD HAVE SOME RIGHTNESS

B. If both could be right, could one view be better or more right than the other? (CIRCLE ONE)

   i. ONE COULD BE MORE RIGHT

   ii. ONE COULD NOT BE MORE RIGHT THAN THE OTHER

Each response was assessed a letter code: A, for Absolutist, if the subject chose response Ai, M, for Multiplist, if he chose response Aii followed by Bii, and E, Evaluativist, if he chose Aii followed by Bi. Each subject was categorized in each domain as an Absolutist, Multiplist or Evaluativist if two or three of the responses in that domain were of the same classification. When a subject had one response from each category in a given domain, he was assigned the classification of Multiplist (M) for that domain. Each subject was then assigned a four letter code, in which the first letter refers to classification in the domain of aesthetic judgments, the second to the domain of value judgments, the third for judgments about truth in the social world and the last for judgments about truth in the physical world.

From the four letter codes, each subject was classified into one of five categories: Absolutist (A), transitioning from Absolutist to Multiplist (T1), Multiplist (M), transitioning from Multiplist to Evaluativist (T2), or Evaluativist (E). Absolutists and Evaluativists had the respective categorization in 3 or 4 of the four domains. Multiplists had the multiplist categorization in 3 or 4 of the four domains or had one A, one E and two Ms. A T1 classification indicates two A's and a T2 classification indicates two E's. This classification is called the Overall Epistemological Understanding (OEU). Additionally, the classification of each subject in judgments about the physical world (PEU) was retained for analysis.

1.2 Assessing Need for Cognition

All administrations of assessment of Need for Cognition reported here used the short form for assessing Need for Cognition, published in Cacioppo, Petty and Kao (1984). The 18-question form for NC was paired with a nine-point Likert scale. The choices ranged from "very strongly agree" to "very strongly disagree". Each of the 18 responses was coded from –4 to +4 so that possible NC scores range from –72 to +72 with higher scores indicated a higher need for cognition. Note that nine of the items are reverse coded. The items from the short form for assessing Need for Cognition appear in Appendix B of this report. The same order of questions was used throughout the studies reported here.

1.3 Assessing Statistical Proficiency

1.3.1 Procedural Fluency

Procedural fluency in statistics is defined in this paper, in Chapter 2, Section 3.1, as the ability both to choose an appropriate statistical process and to complete that process correctly, using proper symbolic syntax. Within the specific domain of hypothesis testing at the introductory level, this aspect of mastery might be tested through an assessment item that instructs the student to perform a hypothesis test with space given to place the four parts of the test. Further, the response area may be specifically labeled: hypotheses, test-statistic, p-value, and conclusion. In order to respond correctly to such an item, the student must be able to write the hypotheses using appropriate symbols, choose the test and corresponding test statistic required by the situation and then calculate correctly the test statistic and associated p-value, all procedural activities. A more

difficult question to assess procedural fluency might be worded more vaguely. For example, "Do the data suggest a difference between the IQ scores of high school graduates and college graduates?" The answer to this question requires the student to perform a hypothesis test, but does not specify such a procedure, thus embedding the procedural decision one layer deeper within the task. In all of the studies reported here, procedural fluency was assessed using "standard" hypothesis test questions. The decision to perform a hypothesis test is obvious, rather than embedded within the problem. The questions that were used are described in more detail in Section 4.3 of this chapter.

1.3.2 Conceptual Understanding

Recall that conceptual understanding in statistics as described in Chapter 1, Section 3.2 is evidenced through the connections that students make between ideas, facts and procedures. A conceptual understanding of the procedures that form hypothesis testing is based on the understanding of sampling distributions and the central limit theorem. The understanding of these ideas, in turn, requires knowledge of the Normal and Student t distributions, at a minimum, as well as a basic understanding of conditional probability.

When one performs a hypothesis test, one is using a sample of the population to make an assertion about some aspect of the population. The reasoning behind inference from a sample depends on knowing the distribution of the sample statistic, often called the sampling distribution. We make a hypothesis about the population and then imagine taking every possible sample of a given size from that population. The sampling distribution is the distribution of the sample statistics that would result from all of the possible samples. We now view the one sample we have collected and its corresponding

sample statistic as one member of a population whose distribution is known. The conclusion of our test is based on the degree to which the value of the test statistic is unusual.

The conclusion that is generated by a hypothesis test is probabilistic, rather than deterministic. One of the two possible conclusions, either the rejection or failure to reject the null hypothesis, is certainly an error. For every hypothesis test, the statistician can calculate the probability associated with both conceivable errors. The p-value associated with the observed test statistic gives the probability of having obtained that value, or one even more extreme, if the null hypothesis is true. If the p-value is small enough, the null hypothesis may be rejected in favor of some alternate hypothesis. However, there is a probability, equal to the value of the p-value, that a sample with a test statistic as unusual or even more unusual than the one we chose would be obtained under the conditions specified by the null hypothesis.

An understanding of the above description is not required for a student to produce the four steps of a hypothesis test. The conceptual understanding, however, is predicted to aid the student in viewing the process as meaningful, rather than as a set of arbitrary steps to be followed. Furthermore, a conceptual understanding of sampling distributions should assist students in predicting how changes in the conditions of an experiment will change the outcome of a hypothesis test. Since conceptual understanding of sampling distributions is not necessarily required for a student to produce a null hypothesis significance test, this understanding must be assessed using items that compel a student to consider how changes in the set up of an experiment will change the outcome of a test. This class of questions includes, for example, “If I were to take a second sample from the population that is twice as big as my first sample and I obtain the same sample mean,

how would the test-statistic (or p-value) of the second sample compare to that of the first sample?"

A more comprehensive list of tasks that demonstrate a conceptual understanding of sampling distribution is given in the following outline:

- Explain what the p-value means in the context of the problem
- Explain the direction of change of the p-value given a change in
  - Test statistic
  - Sample size
  - Null hypothesis value
  - From a one-sided to two-sided test or vice versa
- Discuss critically experimental design features as they relate to
  - Conditions for inference
  - Mathematical structure of inference
- Discuss a test's robustness to conditions for inference.

1.3.3 Strategic Competence

Strategic competence in statistics is described in detail in Chapter 2, Section 3.3 of this work. Two main components of strategic competence of importance to learning about hypothesis testing are, the tendency to recognize that a statistical treatment is called for by a situation and the ability to navigate between the spheres of problem context and statistical knowledge. The recognition that a statistical treatment is called for is characterized by a student's understanding of the need for data, as opposed to anecdotal evidence, to make an informed decision about the world. While this distinction may appear to be a determination about the quality of evidence and fall within the domain of statistical reasoning, recall that strategic competence is closely related to statistical reasoning in this manner and that it is the initial instinct that triggers the need for a particular type of evidence that falls within the strand of strategic competence. An

example of an assessment to target this competency might be to ask a student to explain why the results of a controlled experiment are more conclusive than anecdotal evidence.

The example given above in the procedural fluency section in which the need to perform a hypothesis test is embedded one level down in the problem, "Do the data suggest a difference between the IQ scores of high school graduates and college graduates?" might be considered an assessment of strategic competence. The student must realize that a statistical process is called for to respond correctly to the item. If the item is given, however, as an assessment within a statistics class, there are external cues indicating that a statistical procedure is necessary. It is, therefore, difficult to assess strategic competence within the confines of traditional course assessments because students are cued, by the situation, to give a statistical response. One cannot know whether they would choose a statistical problem solving approach if the problem were not situated within the context of a statistics class. In the studies reported here, strategic competence is measured by having subjects respond to statistical information that is embedded in a contextual task. For further discussion of these tasks, see the description of the "Email from Dad" task in the discussion of the Pilot Study and the "Lyon Diet Study" and "ESP Study" tasks described in the discussion of the Targeted Study.

1.3.4 Statistical Reasoning

Statistical reasoning, as described in Chapter 2, Section 3.4 of this work, is the ability to create and critique statistical claims using the proper standards of evidence and argument of the domain. It is the ability to know what can be inferred from the data and how the context of the experiment mitigates such conclusions. As is the case with conceptual understanding, a student may be able to produce a correct hypothesis test,

including a complete conclusion, though the memorization of a standard format for such conclusions, without an understanding the relevance or scope of the conclusion. In order to assess a student's statistical reasoning skills within this domain, one must ask questions that probe the implications of the results. An example of an assessment to target this competency might ask a student to differentiate between the implications of a controlled experiment versus a survey or to list possible biases that result from a given experimental design and determine how the biases might affect the results of the study. A more complete list of items that assess statistical reasoning are given below:

- Discuss strengths and weaknesses of an experiment and comment on how the experimental design impacts the strength of conclusions that can be made.
- Explain the difference between statistical significance and practical significance.
- Demonstrate understanding of the difference between association and causation and features of experimental design that allow a conclusion of causation.
- Write critically about the results of an experiment.
- Write about the extent to which one believes the results of an experiment.

1.3.5 Assessing Productive Dispositions

The aim of this work is to produce a dispositional attribution model to describe differences in the development of statistical proficiency. An argument was made in Chapters 2 and 3 of this work that Need for Cognition and Epistemological Understanding are two possible dispositions on which the attribution model might be based. Therefore, productive dispositions are assessed in this study using the Epistemological Understanding and Need for Cognition tasks described in Sections 1.1 and 1.2 of this chapter.

2. PILOT STUDY

2.1 Study Design

The purpose of the pilot study was to generate data in order to see if there was a relationship between Epistemological Understanding, Need for Cognition, and the development of the different aspects of statistical proficiency. In addition, the responses to the items on the protocol would be used to refine the items for future studies.

2.2 Participants

All students enrolled in the Introduction to Statistics courses during the Spring 2005 semester were invited via email, sent by either the TA or course instructor, to participate in this pilot study. In addition, I visited each of the classes to encourage more students to participate. Students were told that they would be paid for their time. The research proposal was written and funded to support 50 to 100 subjects. Only 30 students volunteered, representing all four sections of the course. Twenty-three of the subjects were female (77%). There were 11 (37%) freshman in the sample, 14 (47%) sophomores, 3 (10%) juniors and 2 (7%) seniors. There were 11 (37%) Natural Science majors, 7 (23%) students from the College of Liberal Arts, 6 (20%) nursing students, 5 (17%) students from the College of Communication and 1 undeclared major. There were more freshmen than expected in the sample, based on the number in the population. The number of women in the sample and the distribution across the disciplines was roughly what one would expect based on the population.

Half of the subjects reported a cumulative GPA of 3.5 or higher and only 4 reported GPA's lower than 2.5. [1] Twenty-four of the subjects reported having passed

[1]The GPA's given in all studies in the work were self-reported by the subjects. The University at which the

Calculus, with a C or higher, either at the university or high school level. Another 4 subjects reported having taken and passed statistics in high school. The sample was not expected to be a random sample of the population. Due to the nature of self-selection, it is expected that the students who agreed to participate both had positive experiences and had done well in the course. Furthermore, it might be expected that the volunteer subjects would exhibit similar personality traits such as extroversion or agreeableness.

2.3 Protocol

The first page contained questions to collect demographic and mathematical background information. These questions are included in Appendix C. Pages 2 and 3 of the instrument contained the Need for Cognition task. Pages 4 and 5 contained a reading comprehension task designed for preparation for the Graduate Record Examination (GRE) reproduced as Appendix D of this volume. Pages 6 through 9 of the instrument contained the 12-item form for assessing Epistemological Understanding (EU) with three questions per page. Page 10 of the instrument contained a question designed to test basic procedural fluency of hypothesis tests with space below to answer. The statistical table needed to complete the problem was given to the subjects along with the protocol. The table was copied from the textbook used by all of the sections of the course. Pages 11 – 15 of the protocol contained 13 multiple-choice questions designed to test the subjects' conceptual understanding and statistical reasoning about hypothesis testing. Finally, pages 16 through 18 contained a free response task designed to test strategic competence

---

studies were completed has several methods for calculating GPAs and allows students to choose either "credit" or a grade for AP exams. The subjects were not instructed to choose a particular GPA calculation. It is likely that each subject chose the highest possible GPA to report.

and statistical reasoning about hypothesis testing. The statistical tasks can be found in Appendix E, Section 1.

There were three main sources for the multiple choice questions used in the protocol: the ARTIST Comprehensive Assessment of Outcomes in Statistics (CAOS) test, the ARTIST scale for Test of Significance, and the problem bank associated with The Basic Practice of Statistics, Third Edition (Moore, 2003). In most cases the questions were modified from the original version. Sometimes the modification was a change in context. In other cases, the context was preserved but the nature of the question was changed. The remainder of this section explains how the questions on pages 10 through 18 of the protocol were designed to assess statistical proficiency.

2.3.1 Multiple Choice Assessment of Conceptual Understanding

Multiple-choice questions 2, 4, 6, 9, 10 and 12 were designed to assess conceptual understanding of hypothesis testing. These questions can be found in Appendix E, Section 1.2.1. Items 2 and 12 were considered to be a set; both questions probed understanding of the meaning of a p-value, but item 12 gave a context for the hypothesis test while item 2 discussed an abstract hypothesis test. These items were intentionally separated and the order of the options was intentionally different. Question 4 assessed knowledge of the effect of sample size on the results of a hypothesis test. In this item, subjects were not given values from which they could compute a new p-value. They were asked to consider how the conclusions would differ if the same statistic was calculated from samples of different sizes. Questions 6, 9 and 10 assessed understanding of sampling variability. In question 6, subjects were asked about likely outcomes of the sampling process. In question 10, subjects were probed for their understanding of the

effect of sample size on sampling variability. Finally, question 9 probed for understanding about sampling variability by asking about the likelihood of differences in samples means produced by repeated sampling.

2.3.2 Multiple Choice Assessment of Statistical Reasoning

Multiple-choice questions 1, 3, 5, 7, 8, 11, and 13 were designed to test statistical reasoning. These questions can be found in Appendix E, Section 1.2.2. Questions 1, 3 and 13 asked the subject to interpret the meaning of the results of a hypothesis test in the context of a problem. Question 1 assessed whether the subject was aware of the probabilistic nature of the conclusion that is made from a hypothesis test. It asked the subject whether a decision to reject (or not reject) a null hypothesis means the null hypothesis is definitely false (or true), probably false (or true) or neither. Questions 3 and 13 also asked about the strength and scope of a conclusion that can be made from the results of a hypothesis test.

Question 5 was designed to test whether subjects know that an observational study does not allow a researcher to conclude that there is a causal relationship between two associated variables. The situation presented, comparing income and recycling behavior, is one in which causation was a plausible conclusion. I assumed that subjects would find it easy to rationalize the statement: "earning more money allows people to recycle more." This was an intentional part of the item design so that the research would stretch subjects to the limits of their statistical reasoning capabilities.

Items 7, 8 and 13 assessed the knowledge of the scope of a conclusion that can be made when taking into account the experimental design behind the hypothesis test. In item 7, a small sample produces results that are counterintuitive and subjects must choose

between claiming 1) the researcher made a mistake, 2) the sample size is too small to make a conclusion, or 3) the hypothesis tests proves an unbelievable conclusion. In question 8, the subjects chose a response to a student who plans to use a voluntary sample in a hypothesis test, Finally, question 11 asked subjects to mark as valid or invalid conclusions that have been made using the data, which is presented only in graphical format.

2.3.3 Free Response Assessment of Strategic Competence and Statistical Reasoning

I designed the final task of the Pilot Study, the Email from Dad task, to assess strategic competence and statistical reasoning. In the email, "Dad" describes to the subject the results of a hypothesis test of the effectiveness of a blood pressure medication. The context was designed to be realistic, but possibly unfamiliar. In addition to statistical evidence, two pieces of anecdotal evidence about the medication are given. In making a decision about the treatment of Grandma's high blood pressure, the subjects would use their strategic competence and statistical reasoning to formulate a response. The incorporation of the statistical evidence and the anecdotal evidence in a coherent response indicates development of strategic competence. Strategic competence is also demonstrated by the way in which a subject can write about the statistical information within the context of the problem, in this case, the decision about Grandma's medication. Statistical reasoning is assessed based on the comments that are made about the design of the reported medical experiment. Subjects are high in statistical reasoning if they correctly identify elements of the design that support or detract from the strength of the conclusions cited and are able to explain the importance of the design features.

2.4 Procedure

Subjects were tested in groups of 2 to 7 students in a small classroom in the mathematics building on campus. Participants had been told in advance what time to arrive and to expect to work for 60 – 90 minutes. Testing was done during the final exam period of the Spring 2005 semester. Each subject was given the protocol, a formula sheet and statistical tables copied from their textbook. They were told to take as much time as they needed to complete the protocol.

The subjects took between 34 and 105 minutes to complete the protocol, with a median time of 54.5 minutes. The times were roughly symmetrically distributed, with the possibility of a high outlier since the second longest session lasted 83 minutes. The 105 minutes includes moving to another room during the session. It had not been anticipated that any subject would take longer than 90 minutes to complete the protocol and someone else had reserved the room. This subject was one of the four who reported a G.P.A. under 2.5 and one of only 6 who predicted that she would earn lower than a B in the Introduction to Statistics course. These factors may have contributed to the amount of time she spent working on the protocol.

I was in the room during the entire testing session. If students did not have pencils or calculators, they were provided. When students were finished, they turned in the completed protocol to me and were paid for their time.

2.5 Data Analysis

The values of the demographic variables, responses to the multiple-choice questions, NC scores and EU classification were recorded for each subject in a database. The hypothesis test question was scored out of a total of 8 points. One point was

assigned for an attempt at each part of the hypothesis test: 1) hypotheses, 2) reporting of test statistic, 3) reporting of p-value and 4) conclusion. An additional point was assigned for each step if the step had been done correctly. These scores were also recorded in the database. The information in the database was analyzed both quantitatively and qualitatively. Because of the small number of subjects, few quantitative tests could be performed. It was possible, however, through qualitative means, to assess inconsistencies in responses and general patterns of responses.

I transcribed the responses to the "Email from Dad" task, typing them into a word processing program. A Guided Theory approach (Strauss and Corbin, 1998) was used to analyze the transcripts to find commonalities between the responses. Using the categories found in the initial analysis, each transcript was analyzed again and coded based on the categories that resulted from the initial analysis. This process will be discussed in more detail in Chapter 5.

3. POPULATION STUDY

3.1 Study Design

The purpose of this quantitative study was to determine the distribution of Need for Cognition scores and Epistemological Understanding categories among the population of students who take the introduction to statistics course for non-mathematics majors (an algebra-based course) offered by the Department of Mathematics. It was also designed to determine if there is an interaction between these constructs within the target population.

3.2 Participants

All students registered in the four sections of Elementary Statistical Methods in the Fall 2005 semester were eligible to participate in this study. The protocol was administered during the last 10 minutes of class at the end of the second week of classes after approximately 5 total hours of class time in each of four large lecture sections. Of the 434 students registered for the course at that time, 381 submitted surveys, for a response rate of 88%. Where $n \neq 381$ in the analysis, it is the result of an improperly completed survey. A total of 377 participants correctly completed the NC and 376 correctly completed the EU task. Demographic data was not collected from each subject individually, but the following tables give the demographics for the population of the course.

Table 4.1: Population Study Enrollment and Response Rate by section

	Total Enrolled	Respondents	Response Rate	Percent of Anonymous Responses
Section 1	121	105	87%	35%
Section 2	114	98	86%	46%
Section 3	118	108	92%	60%
Section 4	81	70	86%	26%
Total	434	381	88%	43%

Table 4.2: Population Study Percent of students of each sex by section

	Female	Male
Section 1	74%	26%
Section 2	71%	29%
Section 3	80%	20%
Section 4	70%	30%
Total	74%	26%

Table 4.3: Population Study Percent of students in each year by section

	Freshman	Sophomore	Junior	Senior
Section 1	26%	40%	20%	13%
Section 2	14%	36%	30%	20%
Section 3	8%	38%	34%	19%
Section 4	17%	42%	27%	14%
Total	17%	39%	28%	17%

Table 4.4: Population Study Percent of students in each college by section

	Communications	Liberal Arts	Natural Sciences	Nursing	Other
Section 1	4%	27%	43%	19%	7%
Section 2	11%	30%	41%	13%	5%
Section 3	12%	26%	39%	19%	4%
Section 4	14%	22%	40%	11%	14%
Total	10%	27%	41%	16%	7%

3.3 Protocol

Each student was given a packet that contained a cover sheet with space to provide contact information should the student wish to volunteer to participate in future related studies (See Appendix C). Students were instructed that they would be compensated for any future participation and that participation in the current study was voluntary. Pages 2 and 3 of the instrument contained the 18 question short form for NC. Pages 4 through 7 of the instrument contained the 12-item form for assessing Epistemological Understanding (EU) with three questions per page. These tasks were identical to those used in the Pilot Study and were coded in the same manner.

3.4 Procedure

I distributed the protocol packets ten minutes before the end of class during the second week of the semester to students attending the introduction to statistics classes. I had no other connection to the students in the class. Students completed the packets and turned them in to me on their way out of the classroom. The instructor exited the classroom as I passed out the surveys to the students.

3.5 Data Analysis

A score on the Need for Cognition assessment task was calculated for each participant. Subsequently, subjects were classified as Low, Middle or High in Need for Cognition. Those classified as LowNC represent the bottom quartile of participants; those classified as HighNC represent the upper quartile of participants. The MidNC are the two middle quartiles. While many researchers use a median split to classify High and Low NC (see for example, Peracchio and Meyers-Levy, 2005 and Martin, Sherrard and Wentzel, 2005), I chose this classification method in order to better illuminate differences between those high and low in NC. Using a median split would have those quite representative of the center assigned to each of the high and low groups. Even with the classification method I have chosen, those on the boundary of the high and low categories are only about two-thirds of a standard deviation above or below the mean, which does not represent a significant departure from the center. It is, however, an improvement on the median split method of classification. Using a more stringent classification method, for example, setting the boundary at a full standard deviation above the mean, would leave too few participants in that category for any meaningful analysis. The Epistemological Understanding (EU) instrument was coded as specified in Section 1.1 of this chapter. For each subject, both Overall Epistemological Understanding (OEU)

classification and classification in the domain of judgments about the physical world (PEU) were retained.

The data were analyzed using qualitative methods including ANOVA, General Linear Model and Chi-Square procedures to determine the difference by section or choice of anonymity, distribution of the constructs within the population and the relationship between the two constructs.

4. EXAM STUDY

4.1 Study Design

This study had two purposes. Mainly, it would allow me to select subjects for the Targeted Study who demonstrated a basic level of procedural fluency and conceptual understanding of hypothesis testing. Because the first goal was met by embedding questions into the final exam of the four introduction to statistics classes, this study also provided data regarding the development of understanding and misunderstandings by the general population of the students who take the introduction to statistics course.

4.2 Participants

All students who took a final exam in the introduction to statistics classes in the Fall 2005 semester participated in this phase of the study. Those who had not consented earlier in the semester to participate in follow-up studies participated anonymously. Their answers to three multiple choice questions were recorded but their names were not. The responses of those students who consented to participate in future studies were recorded with their names.

A total of 346 students took a final exam in this course. This represents 80% of the number that had been registered at the time of the administration of the Population Study Protocol. Table 4.5 gives the retention rate of students by section. Of the 216 subjects who consented to be contacted for follow-up studies, 166 (77%) took a final exam. The retention rate of these students by section is given in table 4.6.

Table 4.5: Retention Rate of students in Introduction to Statistics

	Enrolled Sept 05	Final Exams	Retention Rate
Section 1	121	98	81%
Section 2	114	81	71%
Section 3	118	105	89%
Section 4	81	62	77%
Total	434	346	80%

Table 4.6: Retention rate of possible subjects in Introduction to Statistics

	Non-anonymous responses	Number taking Final	Retention Rate
Section 1	68	61	90%
Section 2	53	35	66%
Section 3	43	33	77%
Section 4	52	37	71%
Total	216	166	77%

4.3 Protocol

Three multiple choice questions, chosen and designed as a result of the Pilot Study responses, were embedded in the final exams of the four introduction to statistics classes at the end of the Fall 2005 semester. The multiple-choice questions appear in Appendix E, Section 2.1. I designed these questions to test conceptual understanding of

hypothesis testing. In addition, I wrote several long-form hypothesis test questions from which each statistics instructor chose a few to place on the exam. All of the hypothesis test questions given to the instructors appear in Appendix E, Section 2.2. I requested that at least one of the hypothesis test questions include the analysis of a stem plot to assess the reliability of the conclusions that had been made. One of the instructors declined to include the stem plot portion of the question on his exam so his students were eliminated from some of the future analyses.

4.4 Procedure

I approached the statistics instructors about 6 weeks before the end of the semester and asked if they would be willing to include the research items on their final exam. The instructors suggested edits to the items, which I incorporated. With the exception of the decision of one instructor not to include the stem plot portion of a long answer problem, the instructors agreed to embed the questions in the final exams for their classes. I designed the questions based on the text that is used in the course and using my knowledge from having taught and been a Teaching Assistant for the course. Thus, I assumed that the questions would not appear to be out of place to the students taking the exam. In addition, because the responses to the questions would affect the students' grades in the course, I also assumed that they would try to answer the questions correctly. Therefore, the collected responses should reflect the highest quality that the students were able to produce.

4.5 Data Analysis

The data from this study were analyzed using quantitative methods. The the percent of students in the population who chose each response on the multiple choice questions was calculated to find out what students know at the end of the course. I recorded the responses given by all of the non-anonymous students on all of the multiple-choice questions. In addition, I read each response the non-anonymous students gave to the hypothesis test questions and took notes on the quality of the responses. Based on these notes, I assigned a number from 0 to 3 to each subject. A 3 indicated the ability to complete a hypothesis test correctly, without calculation errors, and reason about the reliability of the results of the hypothesis based on a stem plot of the sample data. A score of zero indicated the student's inability to do more than rewrite the problem or attempt a calculation.

I assigned each non-anonymous subject a value, from 0 to 6, representing his demonstration of procedural fluency and conceptual understating of hypothesis testing. The value was the sum of the number of correct responses to the multiple choice questions added to the score earned for producing hypothesis tests. I used an ANVOA procedure with these scores as responses and the NC and EU categorization as factors to determine the existence of a relationship between the development of procedural fluency, conceptual understanding and the constructs from psychology.

5. TARGETED STUDY

5.1 Study Design

The goal of this study was to find differences in the development of those aspects of proficiency that can be related to differences in Need for Cognition and

Epistemological Understanding. In particular, it was designed to find differences in the development of strategic competence and statistical reasoning by students who had already shown proficiency in procedural fluency and conceptual understanding in hypothesis test questions. This study used open-ended interviews to assess differences in the development of strategic competence and statistical reasoning about hypothesis testing by beginning statistics students.

5.2 Participants

Because this study was designed to finding differences in students among students who had already shown a basic level of competence with hypothesis testing, only the students who had correctly responded to two of the three multiple-choice questions that had been embedded on their final exam and had received at least a 2 out of 3 on their completion of hypothesis test items were eligible to participate in this study. The study was also designed to test differences that developed as a result of differences in EU and NC, so only students in either the HighNC or LowNC and Multiplist or Evaluativist categories were originally invited to participate.

In the original study design, I proposed that I would interview at least 4 and as many as 7 students in each of the four categories, HighNC-Multiplist, LowNC-Multiplist, HighNC-Evaluativist, LowNC-Evaluativist. When I counted the eligible subjects, however, there were no more than 7 total available students in any category. Therefore, the Targeted Study was completed with many fewer than expected subjects. Furthermore, as subjects were retested on the EU and NC tasks, a possible test-retest reliability issue began to appear. The test-retest reliability problem will be discussed in more detail later in this work. When the problem occurred in the research, I invited additional subjects to participate. In particular, I extended invitations to students who were originally

categorized as HighNC-Transitional and LowNC-Transitional. A final constraint on subjects was the difficulty in obtaining LowNC volunteers. Given that those low in NC are less likely to enjoy cognitive activities, it is, perhaps, a reasonable conclusion that these students would be less willing to volunteer to be interviewed for an hour about statistics.

Ten subjects participated in the Targeted Study, 8 women and 2 men. There was one freshman in the sample and three students each who were sophomores, juniors and seniors, classified by credit hours. The students were all of a traditional age, ranging from 18 to 21 with a modal age of 19 (four subjects). Nine students reported having a G.P.A. above 3.7, and the tenth subject reported a G.P.A. of 2.89. Nine of the ten subjects reported earning a grade of A in the introduction to statistics course; the tenth subject had earned a C. The subject who earned a C and the subject with the low G.P.A. were not the same subject. Five of the subjects were in fields related to medicine: 2 in nursing, 2 in pre-pharmacy and 1 pre-med. There were two advertising students, and one each in public relations, psychology and Chinese.

The classification of the subjects by EU and NC are given in Tables 4.7 and 4.8. Table 4.7 gives the classification based on the Population Study administration of the protocol and Table 4.8 is based on the Targeted Study administration of the protocol. The difference between the tables illustrates the extent of the test-retest reliability problem associated with the EU and NC instruments.

Table 4.7: Targeted Study Classification of Subjects in September

	Multiplist	Transitionalist	Evaluativist
HighNC	3	2	2
LowNC	1	1	1

Table 4.8: Targeted Study Classifications of Subjects in February

	Multiplist	Transitional	Evaluativist
HighNC	2	1	2
MidNC	1	1	1
LowNC	1	0	1

5.3 Protocol

The protocol for this study consisted largely of an interview designed to assess strategic competence and statistical reasoning of the subjects. In addition, subjects completed a written questionnaire to collect data on grades and to retest Epistemological Understanding and Need for Cognition. Furthermore, subjects provided a written response to the "Email from Dad" task described in the Pilot Study. The demographic questionnaire appears in Appendix C of this work.

The interview portion of the protocol was based on three scenarios of varying degrees of personal relevance. Those low in NC tend to postpone cognitive activity until situational factors affecting cognitive motivation are amplified (Cacioppo, Petty, Feinstein, and Jarvis, 1996). Personal relevance is a factor that creates situational demand; using scenarios of different levels of personal relevance should illuminate this interaction in subjects (ibid). There is literature that relates the tendency of subjects to change opinions and each of EU and NC. Those "higher" in EU are more likely to accept information that is in conflict with their personal beliefs and to change their opinions than those low in EU (Mason and Boscolo, 2004). Those high in NC are more likely to resist counter-attitudinal influence so their attitudes tend to persist over time (Cacioppo, Petty, Feinstein, and Jarvis, 1996). These results indicate that there may be an interaction

between EU, NC and opinion changing. A search of the PsycINFO database for Epistemological Understanding and Need for Cognition did not produce any citations in which both constructs were studied so there is no empirical evidence about the interactions between the constructs and opinion changing-behavior. In particular, it is unknown how a HighNC-Evaluativist will react to evidence that is in conflict with his beliefs because the HighNC subject is predicted by the literature to have a different reaction from an Evaluativist.

Given the importance of relevance and pre-existing opinions, the subjects were questioned over the subject matter of the scenarios, prior to presenting the scenarios. This was done to find out how personally relevant the tasks were and to assess the subject's preexisting opinions within the domain of the tasks. The three scenarios were:

A. Email from Dad – Decision making on medication for Grandma: intended high relevance (university students tend to have aging grandparents so this is a type of decision in which the student may actually be involved in the near future)
B. Statistical evidence for paranormal phenomena – Decision making about opinion on the existence of extra-sensory perception: intended low relevance (not a topic generally considered as one that would be studied statistically; existence of ESP not something that students would have thought a lot about.)
C. Link between diet and heart disease/cancer – Decision making on whether subject would consider a lifestyle change based on the data: intended medium relevance (important generally to subject, but subjects are young enough not to be so concerned with heart disease and/or cancer)

The entire protocol for these tasks, including the preliminary questions, the text read by the subjects and the interview questions designed to assess statistical proficiency can be found in Appendix F. The Email from Dad task was modified from a course assessment used at a liberal arts college in the Northern Midwest (Jordan, 2004). The ESP Study reading task was adapted from a journal article that reported the results of a paranormal psychology experiment (Bem and Honorton, 1994). The Lyon Diet Study reading task was created using, verbatim, the results of a study reported on the website of the American Heart Association (American Heart Association, 2005). The interview questions posed in Appendix F are suggestions. The intent was to have guided, rather than scripted, interviews so that the interviewer would have the leeway to explore further the thoughts and understandings of the subjects.

5.4 Procedure

I met all subjects, except Bradford, on the 10th floor of the mathematics building and conducted the interviews in a conference room. In each case the layout of the room was the same. I sat at the end of a table with the subject on my right. There was a microphone attached to a laptop and a cassette recorder on the table between the subject and me. The subject had access to paper, pens, and a calculator. I had paper for taking notes and the pages of the protocol on the table in front of me. The laptop computer was to my left and I used it to record times and to check that the microphone was picking up the voices. In the case of Bradford, I conducted the interview in my office on the 11th floor of the mathematics building. The layout for Bradford's interview was similar to the others, although the desk was smaller than the conference room table and Bradford was on my left side.

The order of the tasks was the same for all participants except Sarah. I welcomed the subject into the room and asked her to sit in the chair provided. I gave the subject a hard copy of the consent form, which I had previously sent as a .pdf attachment via email. After the subject had time to read the consent form and ask questions, the subject and I signed another copy of the form. After this, I described the general format of the interview. I assured each subject that her statistics instructor would not know what they said or even that they had participated.

The subjects completed the "Email from Dad" writing task first. After that, subjects were given the demographic information, EU and NC tasks. While the subject completed those tasks, I read the written email "response" to prepare for the beginning of the interview phase. The only difference in the procedure for Sarah was that she completed the EU and NC task first, followed by the Email from Dad writing task. After the first interview, I realized that it would be more efficient to switch the order of the tasks. I did not start recording until after the pencil and paper tasks had been completed. I wrote down anything that was said by the subjects prior to the beginning of the recording and typed the utterances as part of the transcript of the session. Once recording began, it continued until the subject left the room.

The interviews began with the preliminary questions designed to assess the relevance of the Email from Dad task. After this, I role-played being the father, calling to follow up on the email. The order of the other two tasks was determined at random. I rolled a die in my office prior to the interview. The Lyon Diet Study was given first if the number was even; the ESP Study task was given first if the number was odd. For both of these tasks, I first asked the preliminary questions and then gave the subject the

article to read. After the subject finished the article, I followed the protocol of questions given in Appendix F.

When the interview was over, I debriefed the subject, explaining the purpose of the study. Subjects were paid $20 for their time and the recordings were stopped after they left the room. The interviews lasted 50 – 70 minutes with most of them finishing in just over an hour.

5.5 Data Analysis

The EU and NC tasks were coded as described in sections 1.1 and 1.2 of this Chapter. The writing samples from the Email from Dad task were grouped with the writing samples on the same task from the Pilot Study. The samples were coded using the categories found in the analysis of the Pilot Study responses. The coded data were analyzed using cluster analysis to ascertain whether differences could be found in the writing samples based on EU and NC categorization.

The transcripts of the interviews were used to create a summary for each subject. The summaries can be found in Appendix G. The categories that had been identified though the open and axial coding of the Email from Dad writing samples were used to identify relevant features of the interviews. In particular, each summary contains a description of the subject's discussion of p-values and experimental design. Further, particular care has been taken to include all types of evidence cited by the participant in each task.

The synopses were the result of a two-phase analysis of the transcripts. Each interview was transcribed in one sitting. Immediately after the interview was transcribed, I wrote a one-page synopsis of my impressions of the interview and subject. Two weeks later, after having completed the analysis of the writing samples, I listened to each

interview tape while reading the transcript and filled in details within the synopsis. I grouped the synopses by EU and NC categorization of the subjects to analyze them for differences in responses among the groups. The summaries were also used to find common themes across the interviews.

Chapter 5: Results

1. PILOT STUDY

This section contains the results of the Pilot Study. Recall that the purposes of the pilot study were, 1) to test for a relationship between Epistemological Understanding, Need for Cognition and the development of the different aspects of statistical proficiency and 2) to select and refine items for future studies. The first part of this section is an analysis of the responses to the multiple choice and true-false questions. The second part provides the results of an investigation of the independence of EU and NC. The third part contains the results of an analysis of the relationship between EU, NC and achievement outcomes. The final part of this section is a qualitative analysis, using grounded theory, of the responses to the Email from Dad task.

1.1 Analysis of Multiple Choice and True-False Items

1.1.1 Statistical Reasoning Questions

Recall that items 1, 3, 5, 7, 8, 11 and 13 comprised the set of items designed to assess statistical reasoning. Question 5, a question about an observational study of the association between income and recycling, was designed to test whether subjects would infer a causal relationship in a context in which such a relationship was plausible but in which the study was not designed to test causality. As expected, subjects did tend to agree that a causal relationship had been found. The sample size used in the example, however, was 1000 so few subjects were tempted to choose the distracter about sample size. This question might produce more variable results if it were rewritten either as a true-false question or with a smaller sample size. Question 8 asked about inferring the

proportion of the population who are opposed to toll roads using, as a sample, letters received by the City Council. The responses to question reinforce the notion that the subjects understand the value of sample size; 10 subjects chose a response that included the large sample size in the justification of a conclusion.

Table 5.1: Pilot Study Results of Statistical Reasoning Multiple Choice Questions

	Q1a	Q1b	Q3	Q5	Q7	Q8
A	10	4	7	18	0	6
B	**17**	**25**	**17**	1	**29**	**20**
C	3	1	5	**11**	1	4
Percent correct	57%	83%	59%	37%	97%	67%

The responses to most of the questions in this set had little variation between subjects. A large number of subjects either answered correctly or incorrectly. When many of subjects gave incorrect answers, they tended to choose the same incorrect response. The results are given in Tables 5.1 and 5.2, the number in bold indicates that the response was the correct response to the problem.

Table 5.2: Pilot Study Results of the Statistical Reasoning True-False Questions

	Q11a	Q11b	Q11c
TRUE	2	**26**	6
FALSE	**29**	4	**24**
Percent correct	97%	87%	80%

In addition to the questions discussed above, Questions 1a and 3 also provide insights into differences in understanding among the subjects. Questions 1a and 1b were about a hypothesis test of defective motherboards. Both questions were written to assess whether a student understands that the decision to accept or reject the null hypothesis might be in error. I was, therefore, surprised that the questions had such different

response patterns. The results for Question 1 suggest that students view the possibility of a type II error, accepting the null hypothesis when it is not true, as more likely than the possibility of a type I error, rejecting the null hypothesis when it is true. A careful read of the problem statement, however, shows that perhaps the students responded to Question 1a as it was written. The question specifies that not all defective motherboards test as defective, giving the possibility of a false positive, but does not mention whether working motherboards might test as defective, the possibility of a false negative. If a subject assumes that there are no false negatives, which is reasonable given the problem statement, then a motherboard that tests as defective is definitely defective and needs to be replaced. Under this assumption, choice A is a reasonable response to the question.

Looking at the individual results on this problem, all 4 students who chose response A for Question 1b also chose response A for Question 1a. So, only 13% of the students in the sample appear to ignore completely the existence of errors of Type I or II. The other 3 subjects who chose option A in Question 1a, chose option B in Question 1b, suggesting that they are aware the possibility of errors but interpreted the statement of the problem to imply that there are no false negatives. Two of the three subjects who responded "none of the above" in Question 1a, gave the correct response in question 1b. Perhaps these students decided not to make an assumption about false negatives.

Question 13 asked students to interpret the results of a hypothesis test about the distribution of M&M colors and to choose the most reasonable conclusion within the context of the problem. Although all of the conditions for inference (as stated in the textbook used in the course) are met, 25 subjects (83%) gave as one of the top two possibilities, that the sample size was too small to draw any conclusion. Half of the subjects suggested the instructor's sample had been unlucky. (With a test statistic more

than 8 standard deviations from the mean of the sampling distribution, she would be quite unlucky indeed!). Seven subjects each were willing to believe that the website was incorrect or thought the machine did not do a good job mixing to create a random sample (the correct answer) and none chose the explanation that the factory was out of green dye on the day the particular bag was produced.

1.1.2 Conceptual Understanding Questions

Questions 2 and 12 probed for understanding of the correct interpretation of the p-value. In both questions, students were asked to identify four interpretations of the p-value as valid or invalid. One of the interpretations was the valid interpretation. The other three interpretations were that the p-value is, 1) the probability of the null hypothesis being true given the data, 2) the probability of replication, and 3) the probability that the results occurred by chance. Question 12 gave the examples within the context of an intelligence experiment and Question 1 was context free, referencing only that a hypothesis test had been done.

Table 5.3 gives the number and percent correct on each of the items in Questions 2 and 12. From the table we see that, in general, students are more proficient in recognizing valid and invalid meanings of a p-value when there is no context. The

Table 5.3: Pilot Study Identification of the Validity of Interpretations of a p-value

Inter-pretation	Correct Interpretation		$P(H_0\|D)$		P(replication)		P(data occurred by chance)	
Situation	In Context	Without Context	In Context	Without Context	In Context	Without Context	In Context	Without Context
number correct	10	27	19	16	23	26	4	11
percent correct	33%	90%	63%	53%	77%	87%	13%	37%

exception is in the case of $P(H_0|D)$. Subjects were more likely to correctly identify this invalid statement as invalid when it was presented in context. Note that the statement that was most poorly identified as invalid, both in context and non-context, was that the p-value is the probability that the data had occurred by chance. The phrase "by chance" is a key word that is used to describe the meaning of the p-value. This interpretation of the p-value may be an initial understanding from which a student can develop a correct understanding. In fact, an interview during the Targeted Study of this work will show that a student who uses the phrase "by chance" does have a reasonable understanding of the meaning of the p-value. The analysis of the free response writings, however, will show incorrect uses of the same phrase.

It would not be appropriate to use a two-sample analysis of the number of correct responses to check for significant differences between context and non-context responses because the responses are not independent. Instead, each subject was categorized in one of four groups for each interpretation of the p-value: correct in both cases, correct in context only, correct in non-context only, or incorrect in both cases. The results appear in Table 5.4. The data show similarities in all cases except the interpretation of the p-value as $P(H_0|D)$. In the other three cases, nearly all of the subjects who made correct identifications in context also made correct identifications without context. This is

Table 5.4: Pilot Study Distribution of Subjects on Assessing Validity of Interpretations of p-values

P-value Interpretation		Correct Interpretation		$P(H_0\|D)$		P(replication)		P(data occurred by chance)	
		Context		Context		Context		Context	
		Correct	Wrong	Correct	Wrong	Correct	Wrong	Correct	Wrong
No Context	Correct	10	17	12	4	22	4	4	7
	Wrong	0	3	7	7	1	3	0	19

not the case for the interpretation $P(H_0|D)$ in which subjects were more evenly distributed among the four possibilities.

A McNemar's Test may be used on data in the form that appears in Table 5.4. The null hypothesis for the test, in this instance, is that the number of subjects in the incorrect-context/correct-without context classification is the same as the number of subjects in the correct-context/incorrect-without context classification. McNemar's Test performed on the outcomes in Table 5.4 found a significant difference for the correct interpretation ($p < .0001$). This means that students who do not have strong understandings or misunderstandings of p-values tend to be better at identifying the correct interpretation of a p-value when it is presented without context.

Question 4 was a two-part question designed to test understanding of how sample size affects sampling distribution, p-value and, eventually, the strength of evidence provided by a test statistic using as context a hypothesis test about the nicotine content of cigarettes. In the first test, one package of cigarettes is used as a sample; in the second, a carton of cigarettes is used to create a larger sample. Eighteen subjects (60%) answered both parts of the question correctly. Another 5 subjects (17%) chose consistent, but incorrect, responses. Four of those 5 people claimed that the p-value would be the same and would provide the same amount of evidence against H_0. Of the 7 subjects who gave inconsistent responses, 3 said that a larger p-value would result in more evidence against H_0, and 2 each said that a smaller p-value would result in less evidence against H_0 and the same p-value would result in more evidence against H_0.

The results from Question 4 show that nearly 70% of the subjects correctly identified that the same test statistic resulting from a larger sample would generate a smaller p-value. This indicates fairly good understanding of the effects of sample size on

the sampling distribution. It is the variance of the sampling distribution that changes based on the sample size with small samples giving rise to higher variation in the sampling distribution. Question 10 also asks how the expected variability in sampling changes based on the sample size. In question 10, the situation is restricted to sampling and is not embedding in a hypothesis test. Only 14 subjects (47%) chose the correct answer to this item, 11 of whom had identified that the larger sample would produce a smaller p-value for the same test statistic. Thirteen subjects (43%) chose the distracter that invoked random sampling, 7 of whom responded correctly to item 4a. These results, which will be replicated in the population study, seem to indicate that students are likely to attend to the key phrase "random samples" before thinking about the effect of sample size on the sampling distribution.

Question 9 was about taking two samples of fish from a pond. Subjects were asked, based on the summary statistics from the first sample, whether the researcher should be surprised about the mean value of the second sample. This question proved to be quite difficult for the subjects. Only 1 answered correctly. 27 subjects (90%) agreed that the results were not unusual because they were within one standard deviation of each other. This result is, perhaps, not surprising given the amount of emphasis that is devoted to the standard normal distribution and in thinking about the fact that unusual occurrences are those that are about two standard deviations from the mean. Students are not attending to the fact that the distribution of the sample mean would be estimated by the sample standard deviation divided by 10 (the square root of the sample size) so it would be 0.5 rather than 5.0. Thus a difference in sample mean of 1.5 is actually a difference of three standard deviations and quite significantly different.

1.1.3 Analysis of the instrument

The results of an analysis of the reliability of the instrument used to assess statistical proficiency in the areas of procedural fluency, conceptual understanding and statistical reasoning were disappointing. The calculated Cronbach's alpha for the 13 items was .08. A principle components analysis did not yield factors that corresponded to the expected three constructs. One possible explanation for this is that the test had too few items (Nunnally, 1967). In a the targeted study, I planned to select students who had developed at least a basic "degenerate model," as discussed in Chapter Two, of statistical proficiency. I would embed questions on the final exams in the statistics courses in order to identify reasonable subjects. It would not be possible to embed more than 13 questions on the exams, so I would be unable to repair the reliability problems with the instrument used in the pilot study.

The reliability score on the statistical reasoning subscale was much lower than that of the conceptual understanding subscale. Therefore, it was decided to assess only procedural fluency and conceptual understanding of potential subjects. Those students who demonstrated, on their final exams, a basic level of procedural fluency and conceptual understanding would comprise the candidate pool for the Targeted Study.

1.1.4 Choosing questions for future research

The conceptual understanding questions that had the highest correlation with both the total score on the 13-item instrument and on the conceptual understanding subscale were questions 4, 10 and 12. These questions were selected to be the basis of the questions that would be embedded on the final exams in the exam study. From previous work with one of the instructors, I knew that she would prefer multiple-choice questions

with five response choices. Question 10 already fit the format, but the other items would need modification.

Question 12 was a difficult question as it was formulated. None of the subjects were able to correctly identify the validity of all four statements. The interpretation of the p-value that contained the phrase "by chance," was the most poorly identified of the statements. Ten subjects chose it as the only valid interpretation of the p-value. It was hypothesized that by removing that interpretation as a choice, the results on this item

Pilot Study Question 12

It is reported that scores on a particular test of historical trivia given to high school students are approximately normally distributed with a mean of 85. Mrs. Rose believes that her 5 classes of high school seniors will score significantly better than the national average on this test. At the end of the semester, Mrs. Rose administers the historical trivia test to her students. The students score an average of 89 on this test. After conducting the appropriate statistical test, Mrs. Rose finds that the *p*-value is .025. Mark each statement as true or false.

A. There is a 2.5% chance that if Mrs. Rose gave another version of the test to her students she would get the same results.

   True False

B. There is a 2.5% likelihood that Mrs. Rose's students' good performance was due to chance.

   True False

C. There is a 2.5% chance that Mrs. Rose's students have the same capabilities as all students across the nation.

   True False

D. There is a 2.5% chance that Mrs. Rose's students would score this high or higher if they actually had the same capabilities of all students across the nation.

   True False

would rise. Further, students would believe that at least one of the remaining definitions was correct. It was, therefore, decided that the choice that the other three interpretations are invalid be removed as a response.

After removing the interpretation illustrated by statement 12b, there were only 7 combinations of responses to the other three parts of the question. They are listed in Table 5.5. After further removing the option to choose that the remaining interpretations are invalid, six options remained. I also removed the response asserting that the p-value represents both P(replication) and $P(D|H_0)$.

Table 5.5: Pilot Study Number of Subjects in each p-value interpretation category.

Interpretation Category	Number of Subjects
All invalid	10
Only $P(H_0\|D)$ valid	7
Only $P(D\|H_0)$ valid	3
Both $P(H_0\|D)$ and $P(D\|H_0)$ valid	3
Only P(replication) valid	3
Both P(replication) and $P(D\|H_0)$ valid	3
All valid	1

The comparison of the results of questions 12 and 2 indicates that students respond differently to questions about the mean of the p-value when interpretations of p-values are embedded in context versus the non-context situation. Because question 12 included context as part of the question, a context had to be developed for the exam item. Since I planned to use question 4 without its "parent question," Question 3, the exam question created from question 12 used the context of question 4, the hypothesis test of nicotine in cigarettes. Question 4 was similarly unsuitable, not being having been written as a five-choice question. An analysis of the responses showed only 6 different choice

pairs. Only one subject chose one choice pair, "Her p-value will be larger and she will find less evidence against H_0," so it was eliminated. The other five choice pairs are represented by the responses given in exam study multiple-choice item 2 in Appendix E. It was expected that this question would be well done by students since 70% of the subjects in the pilot study gave correct responses to it.

1.2 Qualitative Analyses of the Constructs and Learning Outcomes

This section gives the results of qualitative analyses that were designed to find out whether there is a relationship between EU and NC and then whether there is a relationship between the two constructs and GPA. The Pilot Study was designed and funded to have 50 to 100 subjects. Only 30 of the over 400 possible subjects volunteered to participate. As a result, the sample size was smaller than expected. This has implications on both the number of comparisons that could be done and on the strength of the conclusions that could be made from the data. In general, we may use the data to find trends, but will not be able to make predictions for the population.

When one makes multiple comparisons using the same data, as was done here, the use of a Bonferroni correction for the significance level is advised (Weisstein, 2006). Seven analyses are performed on the Pilot Study data. They explore the relationships between, 1) NC and OEU, 2) NC and PEU, 3) NC and GPA, 4) OEU and GPA, 5) PEU and GPA, 6) NC, OEU and GPA and 7) NC, PEU and GPA. Therefore, $\alpha = .1/7 = .014$ will be used to assess significance throughout this section.

1.2.1 Independence of the Constructs

According to a search in PsycInfo, no research has been published that studied both EU and NC. I wanted, therefore, to assess whether the constructs of EU and NC are independent. As was discussed in Section 5.3, the literature on opinion changing

associated with EU and NC indicated that there might be an interaction between the constructs in their relationship with other variables. Because NC scores are quantitative and EU is a categorical variable, I used an ANOVA analysis to test for differences between the means of NC scores between EU classifications. Each subject had been assigned an overall EU score (OEU) and an EU score based on judgments about the physical world (PEU). Because both OEU and PEU will be carried forward to future analyses, the relationship between both EU scores and the NC is considered.

A plot of the NC scores, shown in Figure 5.1, indicates that the distribution of NC scores in the sample is single peaked and symmetric with no outliers. A normal probability plot of the NC scores did not indicate a departure of the data from a roughly normal distribution. When box plots of NC scores by OEU were considered, see Figure 5.2, no outliers appear. The results of the one-way analysis of variance of NC scores with factor, OEU, did not produce significant results; see Figure 5.3. The results, along with

Figure 5.1: Pilot Study Stem and Leaf Plot of NC scores

```
Stem-and-leaf of NC        N  = 30
Leaf Unit = 1.0

    1   -1 6
    2   -1 3
    5   -0 555
    6   -0 4
    9    0 024
   12    0 557
   (7)   1 0001124
   11    1 589
    8    2 134
    5    2 9
    4    3 0
    3    3 8
    2    4 04
```

Figure 5.2: Pilot Study Box plots of NC by OEU: 1 = M, 2 = T2, 3 = E

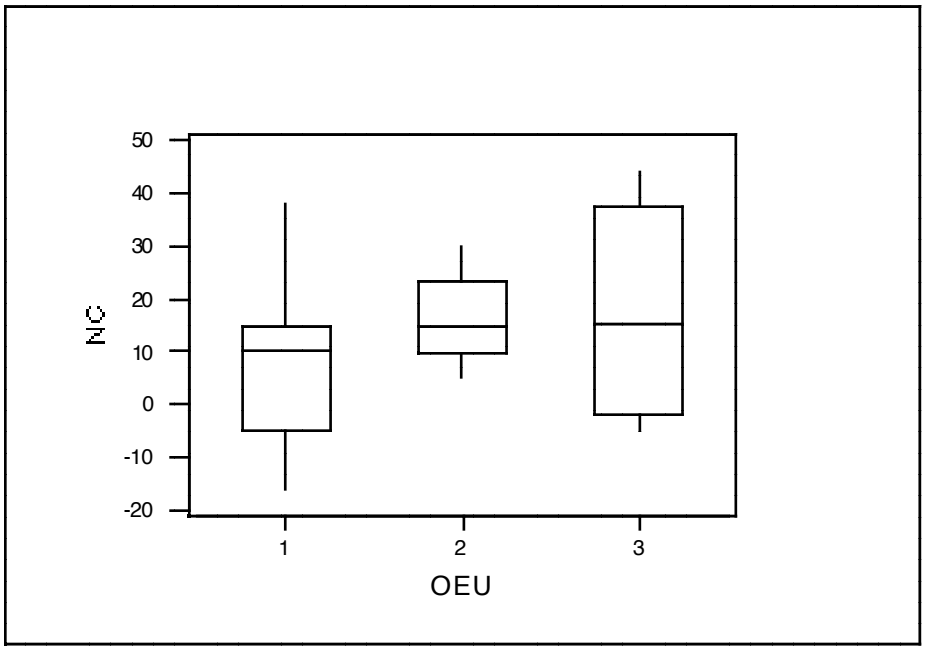


Figure 5.3: Pilot Study ANOVA NC by OEU

One-Way Analysis of Variance

```
Analysis of Variance on NC
Source     DF        SS        MS        F        p
OEU         2       574       287     1.29    0.293
Error      26      5799       223
Total      28      6373

                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
 Level      N      Mean     StDev  ----------+---------+---------+------
     M     15      7.93     14.04  (---------*---------)
     T2     6     16.00      8.90      (---------------*---------------)
     E      8     17.37     19.43          (-------------*------------)
                                   ----------+---------+---------+------
Pooled StDev =    14.93
```

the box plots, however, indicate that the assumption of equal variances between the classes might not be met as the sample standard deviation of the Transitionalists was less than less than half that of the Evaluativists. The total sample size is not large enough to

provide robust results particularly given that the sizes of the groups are different. Therefore, non-parametric analysis of the data was completed as well.

The Kruskal-Wallis test was designed as a non-parametric test that would be used in place of the one-way ANOVA procedure when the assumption of normality is not met (Montgomery, D., 1997). In this test, all subjects are ranked from lowest NC score to

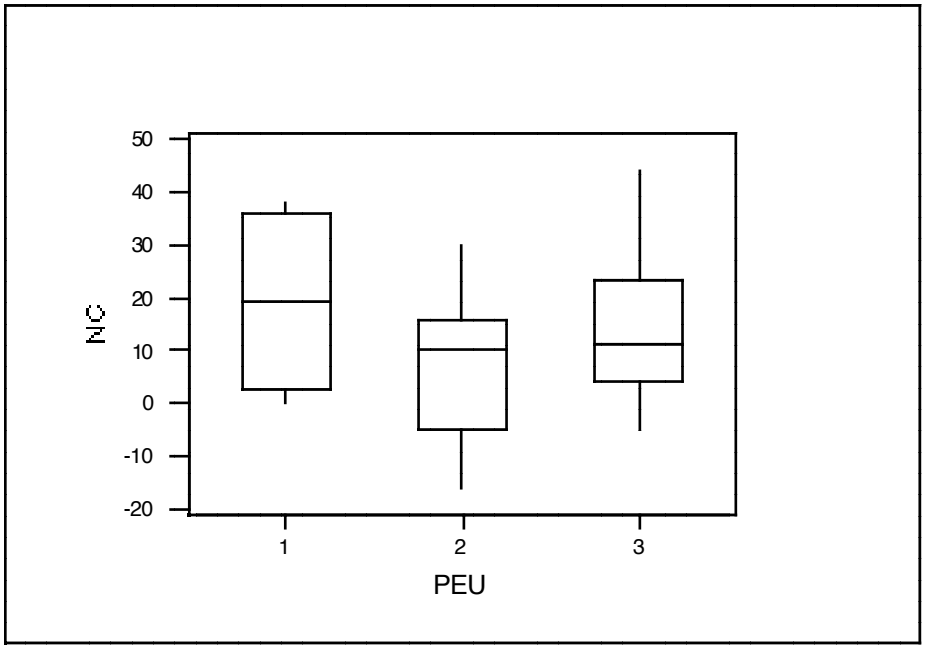
Figure 5.4: Pilot Study Kruskal-Wallis Test of NC and OEU

Kruskal-Wallis Test

LEVEL	NOBS	MEDIAN	AVE. RANK	Z VALUE
1	15	10.00	12.8	-1.46
2	6	14.50	18.1	1.00
3	8	15.50	16.9	0.73
OVERALL	29		15.0	

H = 2.21 d.f. = 2 p = 0.332
H = 2.21 d.f. = 2 p = 0.331 (adjusted for ties)

Figure 5.5: Pilot Study Box Plots of NC by PEU: 1 = A, 2 = M, 3 = E



highest NC score and an analysis is done on the distribution of rankings into the OEU categories. The results of the test are given in Figure 5.4. As with the one-way analysis, no significant results are found, showing no evidence of a relationship between EU and NC.

The same analysis was completed on NC and PEU. The box plots, Figure 5.5, still indicate the possibility that the data do not meet the assumption of normality. A normal plot of the GPAs and the results of an Anderson-Darling normality test (p = .001) confirm that the assumption of normality of the GPAs is not met. Furthermore, the results on the one-way ANOVA, Figure 5.6, indicate that the data do not meet the assumption of equal variances between the groups.

Figure 5.6: Pilot Study ANOVA NC by PEU

One-Way Analysis of Variance

```
Analysis of Variance on NC
Source     DF        SS        MS         F        p
PEU         2       615       308      1.41    0.262
Error      27      5902       219
Total      29      6517
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
 Level      N      Mean    StDev  -+---------+---------+---------+-----
     A      4     19.25    17.35       (--------------*--------------)
     M     14      7.21    13.39  (-------*-------)
     E     12     14.67    15.58          (--------*-------)
                                  -+---------+---------+---------+-----
Pooled StDev =    14.78
```

Figure 5.7: Pilot Study Kruskal-Wallis Test of NC and PEU

Kruskal-Wallis Test

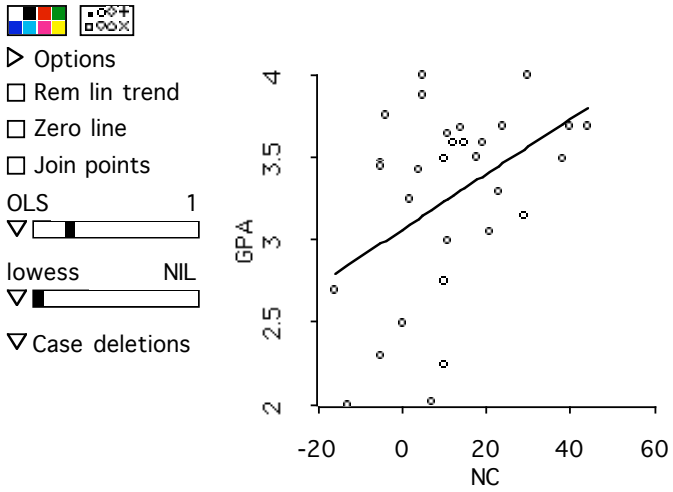
```
LEVEL    NOBS    MEDIAN    AVE. RANK   Z VALUE
    A       4     19.50        18.7      0.79
    M      14     10.00        13.4     -1.23
    E      12     11.00        16.9      0.70
OVERALL    30                  15.5

H = 1.64  d.f. = 2  p = 0.441
H = 1.64  d.f. = 2  p = 0.440 (adjusted for ties)
```

The analysis also finds no significant differences in NC based on PEU classification. The follow-up Kruskal-Wallis test, with results given in Figure 5.7, finds no significant differences in NC scores between the PEU groups. This finding is consistent with the findings for NC and OEU and gives no evidence of a relationship between these constructs.

1.2.1 Relationship between the Constructs and GPA

As was discussed in the experimental design chapter of this work, both EU and NC have been shown to be positively associated with both education level and cognitive capabilities (see for example, Mason and Boscolo, 2004, and Cacioppo et al., 1996). It was therefore hypothesized that there would be a relationship between the three variables, GPA, NC and each of the EU categorizations. In addition to the previously mentioned cautions of sample size and multiple comparisons, the GPA data introduces another issue for analysis. The GPAs of the subjects in the sample are not symmetric; they are left

Figure 5.8: Pilot Study Scatter plot of GPA versus NC



skewed. This is not unexpected given that the sample was a voluntary sample of students asked to participate in a study about their statistics learning. The shape of the distribution of the GPAs adds another caution to the interpretation of the analysis presented here.

The relationship of each of the construct variables with GPA was considered individually. Figure 5.8 shows the scatter plot of GPA versus NC and Figure 5.9 gives the results of the regression analysis of NC on GPA. The p-value for the regression is 0.0154, just over the alpha level chosen for these analyses. The results appear to be consistent with the relationship between NC and GPA given in the literature.

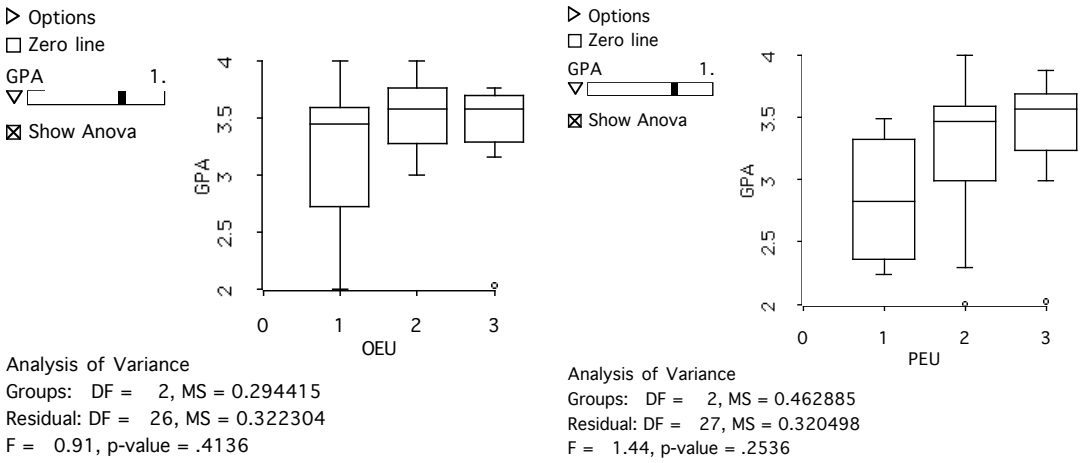
Figure 5.9: Pilot Study Regression Analysis of GPA and NC

```
Data set = Pilot Study, Name of Fit = Coincident
Normal Regression
Kernel mean function = Identity
Response      = GPA
Terms         = (NC)
Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   3.06998         0.122967       24.966     0.0000
NC         0.0167986       0.00651299      2.579     0.0154

R Squared:               0.191978
Sigma hat:               0.525772
Number of cases:              30
Degrees of freedom:           28

Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      1     1.839         1.839       6.65    0.0154
Residual       28   7.74022      0.276436
 Lack of fit   22   5.83444      0.265202       0.83    0.6559
 Pure Error     6   1.90578      0.317631
```

Figure 5.10 shows the box plots for GPA marked by OEU and PEU as well as the results of an ANOVA analysis to test for differences in GPA among the EU groups. While the box plot of GPA marked by PEU suggests that, perhaps, those with an absolutist stance about judgments in the physical world might tend to have lower GPA's than the other groups, neither ANOVA analysis, nor the corresponding Kruskal-Wallis

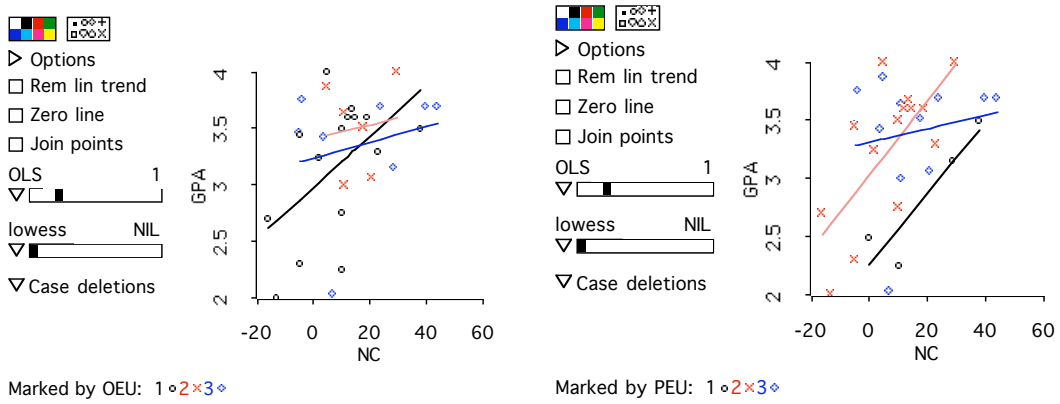
test, produced significant results. The Pilot Study data provides no evidence to suggest a relationship between EU and GPA. Although the box plots suggest that perhaps the condition of equal variances among groups is not met, the sample standard deviations are roughly equivalent, except that the standard deviation of the GPAs for OEU Transitionalists was lower that of the other groups.

Figure 5.10: Pilot Study Box Plots of GPA by EU



Finally, I considered the relationship of both constructs together on GPA. Figure 5.11 shows the scatter plots of GPA versus NC marked by OEU and PEU. The lines shown on the scatter plot represent the results of an ordinary least squares regression analysis of NC on GPA for each EU grouping. The difference in the lines indicates that EU and NC may interact in their effect on GPA. The graph marked by OEU suggests that the relationship between NC and GPA is stronger for Multiplists than for Transitionalists and Evaluativists. The graph marked by PEU suggests that the GPA's of students with an evaluativist perspective on judgments about the physical world are less strongly associated with NC than are those of students in the other classifications.

Figure 5.11: Pilot Study Scatter Plots of GPA versus NC marked by OEU and PEU



To test whether the relationships suggested by the scatter plots were significant, the full models, using NC, factors for EU and the interaction term to predict GPA, were run for both OEU and PEU. The results are given in Figures 5.12 and 5.13. The

Figure 5.12: Pilot Study Regression Analysis Results, NC, OEU, interaction on GPA

```
Data set = Pilot Study, Name of Fit = Unrelated OEU
Normal Regression
Kernel mean function = Identity
Response      = GPA
Terms         = (NC {F}OEU NC*{F}OEU)
Coefficient Estimates
Label          Estimate        Std. Error    t-value    p-value
Constant       2.98508         0.162682       18.349     0.0000
NC             0.0228045       0.0103540       2.202     0.0379
{F}OEU[2]      0.435016        0.516687        0.842     0.4085
{F}OEU[3]      0.263524        0.311784        0.845     0.4067
NC.{F}OEU[2]  -0.0167691       0.0292250      -0.574     0.5717
NC.{F}OEU[3]  -0.0158467       0.0148012      -1.071     0.2954

R Squared:                0.241507
Sigma hat:                0.543848
Number of cases:               29
Degrees of freedom:            23

Summary Analysis of Variance Table
Source         df       SS             MS           F     p-value
Regression      5    2.16601       0.433203       1.46     0.2397
Residual       23    6.80273       0.295771
 Lack of fit   19    5.14016       0.270535       0.65     0.7682
 Pure Error     4    1.66257       0.415642
```

regression using OEU was not significant, but the regression for PEU was significant with p = 0.0117. The problems with the sample and the data cannot be ignored so this result should be used cautiously. It does, however, suggest that a larger study of the relationship between NC, PEU and GPA might be warranted.

Figure 5.13: Pilot Study Regression Analysis Results, NC, PEU, interaction on GPA

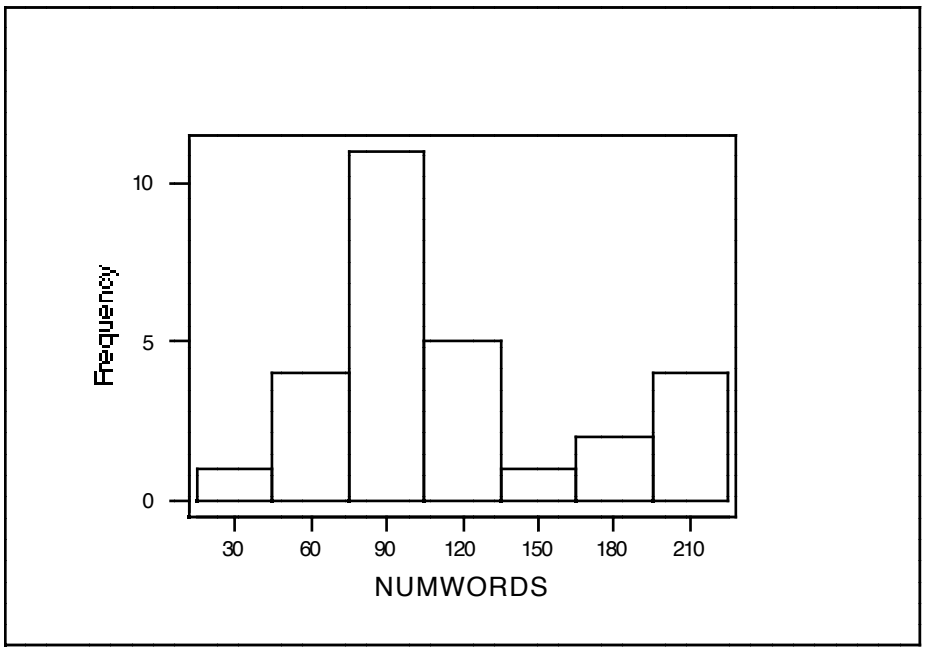
```
Data set = Pilot Study, Name of Fit = unrelated PEU
Normal Regression
Kernel mean function = Identity
Response      = GPA
Terms         = (NC {F}PEU NC*{F}PEU)
Coefficient Estimates
Label          Estimate        Std. Error    t-value    p-value
Constant       2.26615         0.384294        5.897     0.0000
NC             0.0304597       0.0157380       1.935     0.0648
{F}PEU[2]      0.771320        0.410667        1.878     0.0726
{F}PEU[3]      1.05777         0.429347        2.464     0.0213
NC.{F}PEU[2]   0.00137600      0.0185374       0.074     0.9414
NC.{F}PEU[3]  -0.0247273       0.0182064      -1.358     0.1870

R Squared:               0.439793
Sigma hat:               0.472861
Number of cases:              30
Degrees of freedom:           24

Summary Analysis of Variance Table
Source         df       SS             MS           F    p-value
Regression      5   4.21288      0.842575       3.77    0.0117
Residual       24   5.36634      0.223598
 Lack of fit   21   4.22004      0.200954       0.53    0.8397
 Pure Error     3    1.1463        0.3821
```

1.3 Free Response Task – Email from Dad

The "Email from Dad" task provided much response variation. The responses were quite useful in designing future tasks and interview questions. Of the 30 subjects, 2 were unable to do little more than write some numbers and try to do a calculation with the given information, which was not sufficient to reproduce the test statistic and p-value. The other 28 subjects gave responses that were between 27 and 217 words in length, with a mean of 115 words and standard deviation of 49.4 words. As the distribution of words

had a rather interesting shape, it is provided for the reader in Figure 5.14. In addition to the two subjects who could do little more than attempt calculations, 4 subjects either rewrote the values given in the email or used symbols to attempt to recreate the hypothesis that was tested. The other 24 subjects restricted their writing to the email response narrative.

Figure 5.14: Pilot Study Histogram of the number of words written in response to the Email from Dad.



After the Email responses had been typed into a word processing program, they were analyzed using the open coding technique of grounded theory in a line-by-line method to identify the different topics discussed by participants. An organization of the data that resulted from the open coding suggested the existence of 4 broad categories of topics within the responses: 1) decisions about Grandma's medication, 2) discussions of the anecdotes, 3) discussions of the experimental design features, and 4) descriptions of the p-value. The axial coding technique of grounded theory was used to develop a

description of each of the categories. These descriptions are discussed in detail below. In addition to discussions pertaining to the explanation of the doctor's information, 10 of the subjects made personal statements, such as Subject 20, who wrote, "Hope everything works out for grandma and thanks for the check!"

1.4.1 Decisions about Grandma

Fifteen subjects explicitly advised Dad to keep Grandma on the medication. Of the 15, 13 used the study results to justify their decision. One also mentioned that Dad should trust the doctor. Subject 23, who will be discussed in more detail below, advised her father to keep Grandma on the medication as long as the drug was working and was FDA approved. Subject 15 suggested that her father keep Grandma on the medication as long as Grandma's blood pressure did not rise.

Of the 7 subjects who advised that Grandma be taken off the medication, 3 used as evidence a misinterpretation of the study results. The other four claimed that the study results were positive, but that they were mitigated either by the existence of the two anecdotes, the stories of Sally and Larry, or because the sample was too small. One subject who did not make a decision also discussed the mitigation of the positive results of the experiment by the small sample size. There were 4 other subjects who did not give an explicit decision, two were negative about the medication, one was positive and one suggested that her father consult the doctor about medications that have no side effects. Subject 17 suggested that her father put Grandma on a placebo without her knowledge because it has no side effects and, only if that did not lower Grandma's blood pressure, should she be given the medication. The two subjects who could only attempt calculations did not make a decision about the medication.

1.4.2 Discussion of the Anecdotes

Half of the subjects discussed the anecdotal evidence provided by "Larry" and "Sally" in their responses to the email. These responses fell into two major categories: 1) general responses and 2) responses related to the reported study results. Subject 1 gave an example of a general response, writing:

> As for the bad stories you have heard, keep in mind that people are more likely to talk about the negative impacts more than the positive and that you have only heard from two people, think of the thousands of people who did not experience negative effects.

The responses related to the study results also fell into two categories, general comments and those that were specifically linked to the study results. An example of a general comment was Subject 10's claim that he trusted the results of a double blind experiment more than anecdotes from the neighbors. An example of a comment that incorporated the study results and the anecdotes is provided by Subject 6 who told his father that, according to the data, it is unlikely that he would know two people for whom the medication did not work.

Ten of the fifteen subjects who mentioned the anecdotes were not concerned about their existence. They invoked the ideas of sampling variation or lurking variables in their arguments or simply referred to the two people as "outliers." The other five subjects who mentioned the anecdotes expressed concern about the medication based on the stories. Four of those subjects said that the existence of the two anecdotes indicated that more study was needed. The most prevalent justification for the need for more study was that the sample size in the cited study was too small, but one subject indicated that such people should not exist given the study results. He used this as proof that the study results were not reliable.

Subject 30 provided a unique interpretation of the anecdotes following an incorrect interpretation of the study results. She was a senior, pre-pharmacy major, Evaluativist and scored 40 points on the NC task, which, the Population Study results show, placed her in the top quartile for the population of students who take the introduction to statistics under investigation. She expected to earn an A in Statistics and exhibited procedural fluency of hypothesis testing. She wrote:

> Hey Daddy,
>
> What the doctor means, in Layman's terms, is that 0.1% of the time, if they conduct a similar test with the medicine and placebo, you'll get the same or more extreme results. So basically the results they found are a rare situation. I think proof of this is the 2 incidences with Larry and Sally (although Larry's fever could have been cause by other things). I'd take Grandma off the meds – at least until more research is done on it. Love you!
>
> Signed
>
> P.S. Where is this check you have "enclosed" in the email? I'll be watching my mailbox :)

Her interpretation of the p-value as the probability of replication was consistent with her response to Question 12a. The existence of two people for whom the medication did not work confirms, for her, the findings that there is a very small chance that the study could be replicated. This indicates to her that the medication is not useful.

1.4.3 Discussion of Experimental Design Features

There were four categories into which the discussions of the experimental design feature were grouped: 1) definitions of the terms randomized and double blind, 2) assessment of the use of randomization and double-blind on the value of the study results, 3) discussions of the sample size, and 4) comments about the lack of study of side effects. Eleven subjects attempted to define double-blind design and only four of those attempted

to define "randomized." Overall, the definitions given by the subjects were largely correct, although one subject gave the definition of "blind" instead of "double blind." The only error in defining the term "randomized" was a subject who claimed that this meant that subjects were randomly chosen not that they were randomly assigned to treatments. There were two unusual responses. Subject 14 claimed that, since neither the patients nor the experimenters knew what was being given, it could have been "anything" rather than a placebo or medication. Subject 19 expressed skepticism about the existence of the placebo effect.

Ten subjects made a value judgment about the use of the experimental design features of randomization and double blind, with six who both defined and assessed the design elements. Two of the subjects referred to randomization and double blinding very generally, mentioning that these were good design elements, but not explaining why. None of the subjects attempted to explain the specific value of randomization. The subjects who explained the value of double blind did so by saying that it would reduce bias, make the results more reliable and keep the doctors from "making up" their results.

The sample size of 80, with 40 in each treatment group, was retained as part of the design because the textbook used in the statistics course taken by the subjects specifies that a sample size of 40 or more is a reasonably large sample (Moore, 2003). Even if subjects misread the email and thought the sample size was 40 total subjects, according to the material in the text, they should have been satisfied with the size of the sample. Eleven subjects mention sample size and only one of those wrote that the sample was large enough. Some examples of the limitations of the Makemewell study given by the Pilot Study participants based on the small sample size were that the results of the test would not be accurate or reliable or that a test with such a small sample could not be

significant. One subject wrote that the sample size was too small to account for variations in responses to the medication and one wrote that the comparison between the effect size and the sample size made the results of the study less convincing.

1.4.4 Discussion of the p-value

Twenty-three of the subjects referenced the p-value in their email response. When a subject mentioned the value, .001, the response was called a quantitative response. When the value was not specifically mentioned, the response was called a qualitative response. None of the quantitative responses was correct. The qualitative responses were of much higher quality, with a general problem of stating a conclusion to strongly.

The following list gives the different quantitative responses about the p-value. The p-value is the probability that

- the study results are due to chance
- the test results are invalid or wrong
- the null hypothesis is true
- the medication lowers blood pressure
- the medication will not work
- they would obtain these results again
- a member of the sample did not experience a significant drop in blood pressure
- a member of the population would experience a drop in Blood Pressure as large or larger than the average BP drop of the sample
- the drop in BP by a patient taking medication is due to chance

In addition there were two quantitative responses about the p-value that invoked the phrase “by chance”:

- The people in the study who benefited would only have done so less than .1% of the times by chance.
- The p-value is the percent of people whose lack of response to the medication is by chance.

The qualitative responses were better, although some made a stronger conclusion than a hypothesis test warrants. A small p-value:

Reasonable interpretations:

- provides more proof against the null hypothesis
- means you can reject the null hypothesis in favor of the alternate
- provides strong evidence for the alternate hypothesis

Strong interpretation:

- means you can assume the alternate hypothesis is true

Incorrect interpretation:

- means large effect
- indicates that the results are well supported
- indicates effectiveness for the sample tested

Subject 23 discussed the study results in a manner very different from the other subjects. The text of her email is discussed in the next section.

1.4.5 Variation

Subject 23 is a sophomore majoring in Geology. She was an Evaluativist who scored 24 on the NC task, which we will find later is at the upper quartile for this population. Subject 23 expected to earn a B in statistics and exhibited procedural fluency in hypothesis testing but did not do well on the conceptual understanding items. She wrote:

Hey Dad,

Thanks for the check. I needed to go to the grocery store and it really helped me out.

As for Grandma, I hope she's feeling well. How has the medicine done? Does it lower her blood pressure? Is she feeling healthy? These are the two most important questions. The experiment the doctor described sounds good. I can't find anything wrong with it, but I hope the FDA has approved the drug. If they have then I trust them enough to have Grandma stay on the medicine. The best way of knowing whether you should keep her on Makemewell is to talk to her and find out how she is feeling. None of the statistics matter because she is an individual and the statistics are only a good measure of an entire population. Give mom and grandma my love. I'll be home soon.

Love,

"kiddo"

This subject was the only subject to recognize that the statistics are general measures used to describe trends in the population, but the effect on an individual is specific to that individual and not necessarily part of a trend. The writing represents a sophisticated understanding of variation and applications of statistics in the real world.

Only four other subjects mentioned variation at all. Subject 19 stated that when an average drop is reported one cannot assume that everyone's reaction is the same, but did not use that information in any way in her argument. Subjects 1, 11, and 22 all used a definition of sampling variation to explain why their father should not be worried about Sally and/or Larry. Subject 21 specifically linked Sally and Larry to sampling error, claiming that "they could be part of the few that the drug did not work well on." Subject 11 told his father not to worry about Sally because she is probably "one of those blood-pressure-raised-by-chance people" in which case, "it's nothing to worry about since there are probably thousands of people on the medication and some small percent's blood pressure will rise by chance alone." Subject 1 made a similar argument telling his father,

"it is always expected that a few people will vary from the average" and "it is merely coincidence" that he happens to "know two of them."

The results of the analyses of the writing samples were used to design future studies as well as to illuminate differences in understandings students exhibit in context. The mention of side effects had not been anticipated prior to the completion of the pilot study. Because of the number of subjects who did ask about side effects in their email responses, an answer to those queries was prepared for the clinical interviews completed in the targeted study.

As discussed previously, open coding of the written responses was used to identify categories within the data. The open coding was done using a line-by-line analysis (Corbin and Strauss, 1998). The following eight categories emerged from the open coding progress to be used in the future analysis of the writing:

- Decisions about the medication
- Definition of Randomized and Double-Blind
- Assessment of Randomized and Double-Blind
- Discussion of Sample Size
- Discussion of Side effects
- Discussion of Anecdotes
- Discussion of p-value/study results
- Personal Statements

The particular coding used within the categories, developed through the use of axial coding, will be discussed in more detail in the results about the targeted study.

2. POPULATION STUDY

This section describes the results of a large sample study of EU and NC in the population of students taking an algebra-based introduction to statistics course. The sample in the Pilot Study was a voluntary sample rather than a random sample. Since EU measures the amount of cognitive activity people enjoy engaging in, it seems reasonable to assume that those high in NC would be more likely to volunteer to participate in a study about statistics learning than those low in NC. For this reason, one cannot assume that the distribution of NC scores among the pilot study participants is representative. In order to classify subjects as globally high (or low) in NC, a study of the distribution of the construct within the population of students was warranted. Such a population study also allowed me to find the distribution of NC categorization within the population and test again the independence of the two constructs.

The data collected in the Population Study is also used in the Exam Study. Nearly 20 quantitative analyses are completed with this data in the search for possible relationships between EU, NC and learning outcomes in statistics. Using a Bonferroni adjustment and a base significance level of $\alpha = 0.1$, the results of any analysis performed here should be less than 0.005 in order to attain a level of significance.

2.1 Removing Unnecessary Factors

The EU and NC instruments were administered to students in four sections of the course. Some students agreed to be contacted for further participation so there were both anonymous and non-anonymous respondents. In order to investigate differences in response types based on section and choice of anonymity, a general linear model on NC, using the factors: section, choice of anonymity and PEU and all two- and three-way interactions was run. The results of this analysis, given in figure 5.15 below, indicated, at

$\alpha = .1$, no main effects or interactions involving section or anonymity. Due to empty cells a full model incorporating OEU could not be run. A main effects model, however, on

Figure 5.15: Population Study Results of GLM; full model: Class, namecode, PEU and interactions

```
Factor   Levels Values
PEU           3     A     M     E
Class         4     1     2     3     4
Namecode      2     0     1


Analysis of Variance for NC
Source         DF     Seq SS     Adj SS     Adj MS       F      P
PEU             2     2208.3     1859.1      929.6    2.88  0.058
Class           3      839.0      302.4      100.8    0.31  0.817
Namecode        1       14.8       20.4       20.4    0.06  0.802
PEU*Class       6     2298.6     2016.9      336.1    1.04  0.398
PEU*Namecode    2     1413.0     1477.3      738.6    2.29  0.103
Class*Namecode  3     1810.3     1581.5      527.2    1.63  0.181
PEU*Class*Namecode
                6     1966.0     1966.0      327.7    1.01  0.415
Error         354   114306.3   114306.3      322.9
Total         377   124856.2
```

Figure 5.16: Populations Study Results of GLM: Main effect model: Class, namecode, OEU

```
Factor   Levels Values
OEU           5     A     T1    M     T2     E
classcod      4     1     2     3     4
anoncode      2     0     1

Analysis of Variance for NC

Source      DF     Seq SS     Adj SS     Adj MS       F      P
OEU          4     3948.2     4063.7     1015.9    3.12  0.015
classcod     3      854.2      858.2      286.1    0.88  0.452
anoncode     1       21.8       21.8       21.8    0.07  0.796
Error      368   119865.6   119865.6      325.7
Total      376   124689.8
```

NC with factors: section, choice of anonymity and OEU indicated no effects for section or choice of anonymity. Those results are given in figure 5.16. Furthermore, chi square analyses did not provide evidence against the hypotheses that both OEU and section and

OEU and choice of anonymity are independent (for OEU and section, chi-square = 5.86, df = 9, p > .75; for OEU and anonymity, chi-square = 3.45, df = 3, p > .30). Therefore, in all future analyses the data will be collapsed across section and choice of anonymity.

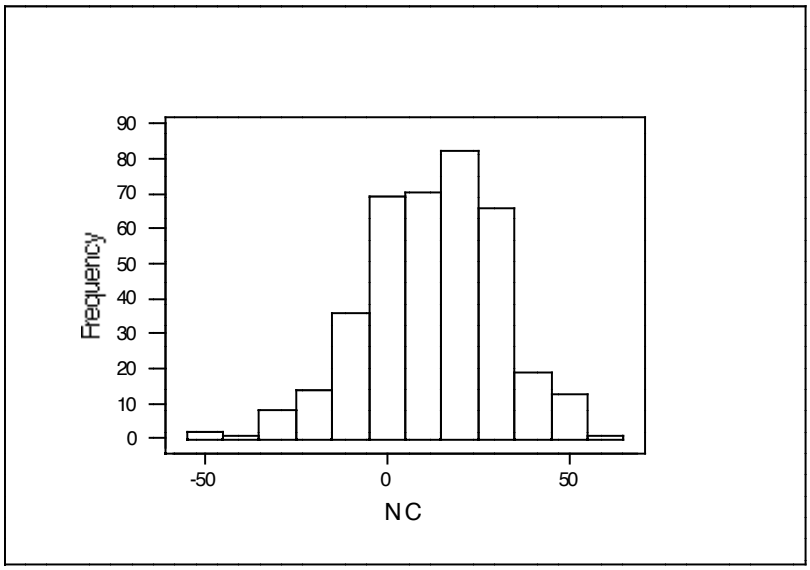
2.2 Need for Cognition Results

The scores of the population on the Need for Cognition scale ranged from –53 to 59. They were roughly symmetrically distributed with a mean of approximately 12 and standard deviation 18. Figure 5.17 below gives the summary statistics and Figure 5.18 is a histogram of the NCS scores. This study replicated findings on reliability and factor

Figure 5.17: Population Study Descriptive Statistics NC

Variable	N	Mean	Median	StDev	Min	Max	Q1	Q3
NC	381	11.79	13.00	18.157	-53	59	-0.5	25

Figure 5.18: Population Study Histogram of Need for Cognition Scores



structure of NCS from previously published works (for a meta-analysis of NC studies, see Cacioppo, Petty, Feinstein and Jarvis, 1996). The calculated Cronbach's alpha was .88.

A principle components analysis found that the first component accounts for 34.4% of the variation, the second and third 8% and 6% respectively. These figures are similar to those cited in Cacioppo, et al. (1996) in which the authors claim that this is evidence of a single factor structure. The factor loadings of the first factor create a weighted average with most weights being roughly .2, .25 or .3.

2.3 Epistemological Understanding Results

There is general consensus in the literature that subjects' epistemological understanding can be classified into three categories (Krettenauer, 2004). Deanna Kuhn calls the categories: Absolutist, Multiplist and Evaluativist. Other researchers working in the domain use different tasks and different coding methods to classify subjects according to their epistemological understanding. In addition, there appears to be a large amount of variation in the percent of subjects classified in each category. Weinstock (2005) classified only 9% of subjects with some college to have transitioned into evaluativist understanding. The percent of Evaluativists among his college graduates was 28%. In contrast, over 40% of the adolescents studied by Krettenauer (2004) were identified as Evaluativists. Full classification results from the literature are contained in Table 5.6.

The protocol and classification used in this study most closely resemble that used in Kuhn, Cheney and Weinstock (2000). Those researchers report the classification results of 107 subjects of roughly equally sized groups of 5th graders, 8th graders, 12th graders, undergraduates, community college students and professionals. The classification results are quite similar to those found in this study. In both cases, three

percent of the subjects studied were classified as not yet exhibiting the Overall Epistemological Understanding (OEU) of Multiplist. In this study, 43% of the subjects

Table 5.6: Population Study Percent of subjects by OEU classification

	Absolutist	Transition 1	Multiplist	Transition 2	Evaluativist
Kaplan (2006)	3%		43%	31%	26%
Kuhn et al. (2000)	3%		58%	13%	23%
Weinstock (2005)					
High School	63%		34%		3%
Some College	48%		43%		9%
College Graduate	28%		44%		28%
Some Graduate	33%		33%		33%
Krettenauer (2004)	24.5%		21%		40.5%

were classified as Multiplists (M), 31% as transitioning between Multiplist and Evaluativist (T2) and 26% as Evaluativists (E). Kuhn and her colleagues found that 58% of the subjects in their sample were Multiplists, 13% transitioning between Multiplist and Evaluativist and 23% as Evaluativist. The population tested in this study was comprised

Table 5.7: Percent of subjects by PEU classification

	Absolutist	Multiplist	Evaluativist
Kaplan (2006)	20%	31%	49%
Kuhn et al. (2000)			
Grade 5	30%	50%	20%
Grade 8	12%	54%	32%
Grade 12	19%	43%	38%
Undergraduate	10%	50%	40%
Community College	10%	45%	45%
Professional	22%	34%	44%

only of undergraduates so it may be reasonable to expect that they would demonstrate a higher level of EU classification. Therefore, these results are not surprising in comparison to those of Kuhn et al. (2000). When classified only by judgments about the physical world (PEU), 20% of the subjects in the current study held absolutist views in this

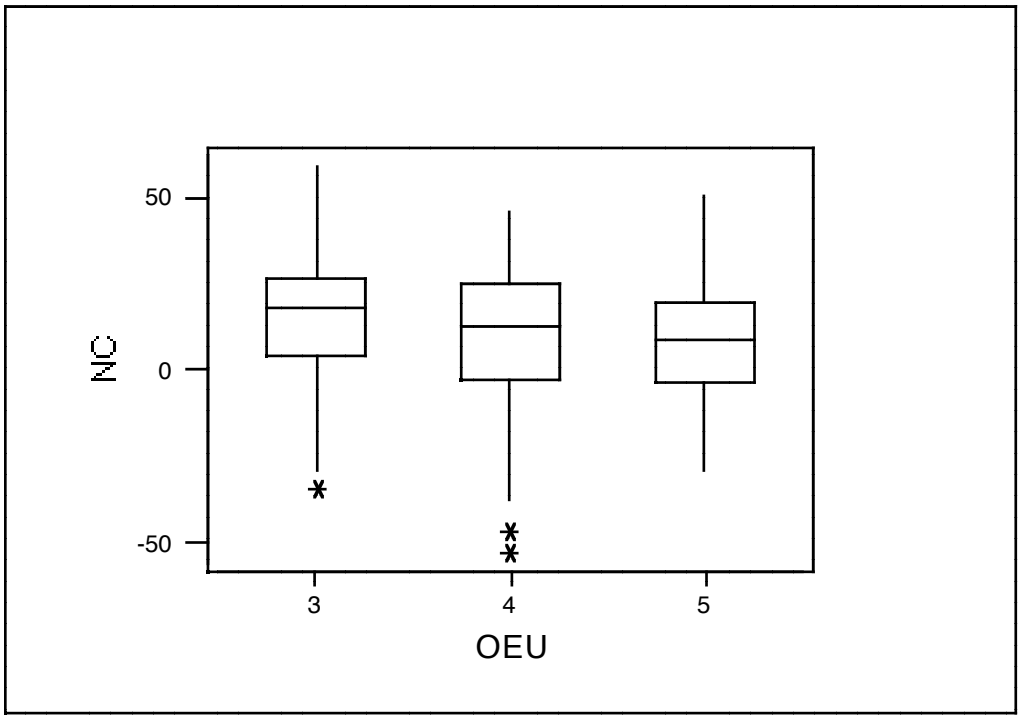
domain, 31% were Multiplists and 49% were Evaluativists. While the results of Kuhn et al. (2000) suggest that one should expect fewer undergraduates in the Absolutist category and more in the Multiplist category (see table 6 for full results), the findings of this study seem to be generally in keeping with previous results.

2.4. Relationship between NC and EU

The secondary goal of the Population Study was to assess the independence of the two constructs, Need for Cognition (NC) and Epistemological Understanding (EU). To test this a one-way analysis of variance was run on NC using, first, Overall Epistemological Understanding (OEU) and then, Epistemological Understanding in judgments about the physical world (PEU) as factors. Because of the small number of subjects who exhibited an OEU classification below that of Multiplist, those subjects were eliminated from the analysis. While both analyses produced small p-values, (NC by OEU, $F = 4.90$, $p = 0.008$; NC by PEU, $F = 3.38$, $p = 0.035$), using the Bonferroni correction, neither achieved the level of significance specified for this analysis. Because the p-value associated with test of NC and OEU was quite low, the relationship was investigated further.

The results of a Tukey pair wise comparison with $\alpha = .05$ (family rate) suggest a difference only between Multiplists and Evaluativists. The Multiplist group appears to score higher on NC than do the Evaluativists. The effect sizes are small, with the mean NC of the Multiplist group less than a half of a standard deviation above the mean NC of the Evaluativist group. The variability of the sampling distribution of a test statistic becomes smaller as the sample size becomes larger. For large samples, such as that used in this study, an observed test statistic that is not very far from the hypothesized mean of the sampling distribution may produce a significant p-value. Given the small effect sizes,

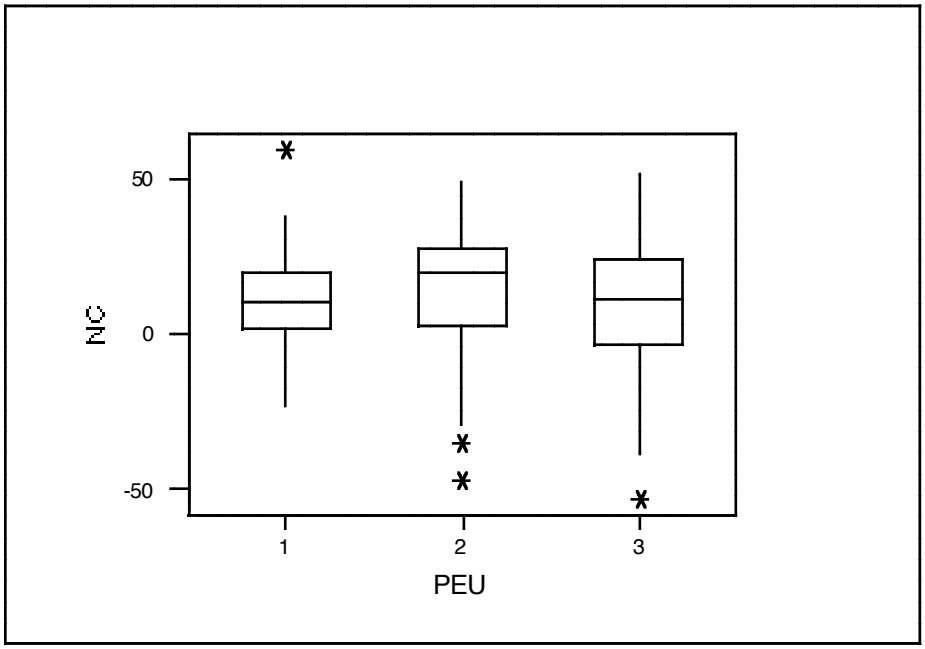
Figure 5.19: Population Study Box Plots of NC by OEU: 3 = M, 4 = T2, 5 = E



it may be that these significant differences are an artifact of the relatively large sample size rather than a clear difference in NC scores across EU categorizations. Furthermore, they may not be practically significant.

Previous studies of EU have shown that EU is associated with educational level. Weinstock and Cronin (2003) claim that there is evidence that educational level and not age is responsible for the developmental differences in EU evidenced by various populations. In particular, the results found by Weinstock (2005) indicate that the percent of people who have a college degree that may be classified as Evaluativists is 28%, a significant rise over the 9% of people who have completed some college coursework and are categorized as evaluativist in their epistemological understanding. This suggests that undergraduate students are undergoing development in epistemological understanding. The relationship found between NC and EU classification may indicate

Figure 5.20: Population Study Box Plots of NC by PEU; 1 = A, 2 = M, 3 = E;



that NC is a mediating factor in EU development. Using the initial results as an indicator, it appears as though those with high NC are less likely to have transitioned from a Multiplist level of EU to an Evaluativist level of EU.

In order to test this hypothesis, the subjects in this study were divided into three groups: HighNC, MidNC and LowNC. The HighNC group comprises the top quartile of NC scores and the low NC is the bottom quartile. The MidNC, therefore, represents the middle two quartiles of the data. There were 99 subjects in the HighNC group with scores ranging from 25 to 59, the mean was 33 and standard deviation was 7.5. The 95 subjects in the LowNC group had scores ranging from –53 to –1, with a mean of –12 and standard deviation of 10.4. The 187 subjects in the MidNC category had a mean of 12.5 with standard deviation equal to 7.3 with scores ranging from 0 to 24. Tables 5.8 and 5.9 show the number of subjects in each NC classification by OEU and PEU classifications, respectively.

Table 5.8: Population Study Number of Subjects by NC classification and OEU

	Multiplist	Transition2	Evaluativist	Total
LowNC	29	31	30	90
MidNC	75	51	50	176
HighNC	52	33	14	99
Total	156	115	94	365

Table 5.9: Population Study Number of Subjects by NC classification and PEU

	Absolutist	Multiplist	Evaluativist	Total
LowNC	15	23	56	94
MidNC	44	55	86	185
HighNC	16	40	43	99
Total	75	118	185	378

Chi-square analyses of NC and both OEU and PEU on the tables produce similar to the results of the ANOVA analysis (OEU, chi-square = 12.93, d.f. = 4, p = .0116; PEU, chi-square = 10.03, d.f. = 4, p = .04). It is more likely that there is a relationship between OEU and NC than between PEU and NC, but given the number of analyses being done with this data, neither approaches the level of significance being used in this study.

The most significant contribution to the test statistic in the analysis of NC and OEU was the HighNC-Evaluativist cell. In particular, many fewer subjects who exhibit high NC than expected are Evaluativists. If this result is found, in the future, to be significant, it may be that those who value cognitive thought and experiences take longer to adapt their views of the world. In other words, it may take more evidence to convince them to change their opinion or outlook.

The relationship between EU and NC has previously not been studied explicitly. Findings from studies of the individual constructs may predict the interaction between the two constructs. Mason and Boscolo (2004) found that subjects high in EU were more

likely to accept information in conflict with their personal beliefs and change their opinions as a result of the new information. Cacioppo et al. (1996) report that subjects high in NC are more likely to resist counter-attitudinal influences so their attitudes are persistent over time.

3. EXAM STUDY

The section reports the responses given by students on assessment items that were embedded in the Fall 2005 final exams of the four introduction to statistics classes. Three hundred forty-six students took a final exam in an introduction to statistics course at the end of the Fall 2005 semester. Each exam had three common multiple choice questions and several different hypothesis test questions embedded within them. The responses by all 346 students on the embedded multiple-choice questions were recorded. The responses to the multiple-choice questions are analyzed in the next section. The following section is an analysis of the responses given by the subset of students who agreed to participate in future studies.

3.1 Overall Results on the Multiple-Choice Questions

The first multiple-choice question required students to select the correct interpretation of a p-value in context. The results are given in Table 5.10. More than half of the students were able to identify the correct interpretation of the p-value when the question was asked in this limited format. Recall that the item in the Pilot Study from which this question was derived included an additional interpretation of the p-value, that is, the p-value is the probability the results occurred by chance. Furthermore, there were two other response combinations that had been chosen by subjects in the Pilot Study but did not appear in this question. Comparing the results here to those from the Pilot Study,

it appears as though the students who may have been distracted by the "by chance" phrase and the two missing response categories tended, in larger numbers, to choose the correct answer. One should note, however, that more than one-third of the students either believe that the p-value is the probability of the opposite conditional statement or do not understand that there is a difference between the two conditionals. A chi-squared analysis of the responses to this item by class indicated no significant difference ($X^2_{6\ d.f.} = 19.08$, $p = .0867$) between the classes.

Table 5.10: Exam Study results for multiple choice question on p-value interpretations

	Percent Choosing	
Response	pilot study	exam study
All given interpretations invalid	33%	NA
Only $P(H_0\|D)$ valid	23%	22%
Only $P(D\|H_0)$ valid	10%	52%
Both $P(H_0\|D)$ and $P(D\|H_0)$ valid	10%	13%
Only P(replication) valid	10%	8%
Both P(replication) and $P(D\|H_0)$ valid	10%	NA
All valid	4%	5%

The second multiple-choice question assessed students' understanding the effect of the sample size on the sampling distribution. This question was quite well done by students: 70% answered correctly. The conditions for a chi-squared analysis to test for independence of response by section were not met because more than 20% of the expected counts were less than 5. An analysis by section of number correct versus number incorrect was not significant ($X^2_{3\ d.f.} = 6.93$, $p = .0741$). The third multiple-choice question also tested understanding of the effects of sample size on the sampling distribution, in particular, on the variation of the sampling distribution. This question

was about a sample and was not embedded in the context of a hypothesis test. In the pilot study, the response: both samples would have the same chance of having an unusual number of brown candies because they are both random samples, was selected at the same rate as the correct response, that the smaller sample is more likely to be unusual. This result was replicated in the exam study, with 43% choosing the correct response and 42% choosing the random sample distracter. Twelve percent thought there is more variability in a proportion among larger samples. There was no difference among sections on the responses to this item.

3.2 Interactions among EU, NC and learning outcomes

Of the 216 students who had agreed to be part of future studies, 164 took a final exam in the course. Since these subjects were not anonymous, their responses to the embedded questions could be analyzed for effects as a result of EU and NC classification. In addition, the 75% retention rate exhibited by this subgroup of the students gave rise to questions of whether EU and/or NC affect the retention of students in an introduction to statistics class.

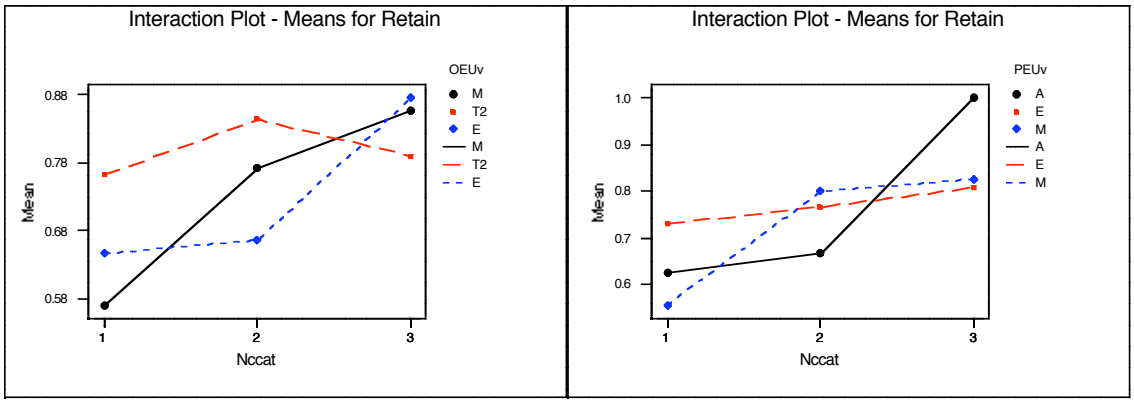
3.2.1 Retention Rate

Recall from Chapter 4 that the overall retention rate based on the number of students who took the final compared to the number of students who had been registered for the class when the population study was completed was 80%, with retention rates in each section ranging from 71% to 89%. The retention rate of non-anonymous subjects had more variability, ranging from 66% to 90% for the four sections.

Figure 5.21 shows the interactions plots for the percent retained by NC category and OEU and PEU categorization, respectively. If the piecewise lines in an interactions

plot are roughly piecewise parallel, it indicates no interaction between the constructs on the outcome variable. If, as is the case in these plots, the lines cross, an interaction is

Figure 5.21: Exam Study Interactions Plot for NC, EU and retention



suggested. From Figure 5.21 it appears as if retention rate is roughly positively associated with NC except in the case of the HighNC-Transitionalists who are retained at a lower than expected rate. Figure 5.21 also confirms the possibility of a main effect for NC, but indicates that HighNC-Absolutists are retained at a higher than expected rate while LowNC-Multiplists are retained at a lower than expected rate.

Retention is a binary value; a 0 is assigned to a student who did not finish the course and a 1 is assigned to students who completed the course. Linear regression models are inadequate for analyzing binary responses "because they can permit probabilities less than zero and greater than one" (Ramsey and Schafer, 2002, pg. 579). The most useful alternative in this case is logistic regression (ibid). In order to analyze the effect of EU and NC on Retention, factors for Section, OEU, PEU and NC were created from the data. In addition, the interactions between the factors for NC and each of the OEU and PEU factors were created.

The results of a logistic regression on Retention using factors Section, NC, OEU and the interaction between OEU and NC are given in Figure 5.22. Based on the p-

values, the coefficients for all three of the section factors are significant as is the coefficient for the NC category of Evaluativist. There is no need in logistic regression to

Figure 5.22: Exam Study Logistic Regression Results, Full Model using OEU

```
Data set = NonanonresultsFULL, Name of Fit = Retain-OEU
Binomial Regression
Kernel mean function = Logistic
Response      = RETAIN
Terms         = ({F}SECTION {F}NCCAT {F}OEU {F}NCCAT*{F}OEU)
Trials        = Ones

Coefficient Estimates
Label                  Estimate        Std. Error     Est/SE    p-value
Constant                1.07735        0.665487        1.619     0.1055
{F}SECTION[2]          -1.59342        0.528149       -3.017     0.0026
{F}SECTION[3]          -1.02904        0.557663       -1.845     0.0650
{F}SECTION[4]          -1.44478        0.516073       -2.800     0.0051
{F}NCCAT[2]             1.09311        0.706362        1.548     0.1217
{F}NCCAT[3]             1.79035        0.813038        2.202     0.0277
{F}OEU[2]               1.02498        0.796548        1.287     0.1982
{F}OEU[3]               0.738403       0.798780        0.924     0.3553
{F}NCCAT[2].{F}OEU[2] -0.399763        1.02211        -0.391     0.6957
{F}NCCAT[2].{F}OEU[3] -1.18845         0.989324       -1.201     0.2296
{F}NCCAT[3].{F}OEU[2] -1.25431         1.12710        -1.113     0.2658
{F}NCCAT[3].{F}OEU[3] -0.860406        1.46413        -0.588     0.5568

Scale factor:                  1.
Number of cases:             216
Number of cases used:        210
Degrees of freedom:          198
Pearson X2:              227.271
Deviance:                208.790
```

check that the data have constant variance and are without outliers as those are not conditions for logistic regression. In addition, residual plots are of little use because the outcome variable has only two values (Ramsey and Schafer, 2002).

Ramsey and Schafer suggest that model checking be done in the form of informal testing of extra model terms. Because the model was suggested by the interaction plots, rather than adding extra terms, sub-models of the full model were examined instead. In order to assess the value of sub-models, one may calculate Akaike's information criterion (AIC), a measure of the relative information given by a sub-model when compared to the

full model (Cook and Weisberg, 1999). The most efficient model is the one with the minimum value for AIC. Using both backward and forward selection methods of sub-model examination suggested that the best sub-model for predicting Retention, when OEU was the EU classification used, is the one containing only the factors for Section and NC. The same results were found when the EU classification PEU was using in the full model.

The results of the regression analysis on Retention using the factors Section and NC are given in Figure 5.23. The coefficients that result from a logistic regression do not have the same interpretation as the coefficients in a linear regression model. In order to interpret the logistic coefficient, one must calculate the value $e^{(\text{coefficient})}$. The resulting value tells you how much more likely a person in that category is, in this case, to have

Figure 5.23: Exam Study Results of Logistic Regression - Submodel for Retention

```
Data set = NonanonresultsFULL, Name of Fit = Retain
Binomial Regression
Kernel mean function = Logistic
Response      = RETAIN
Terms         = ({F}SECTION {F}NCCAT)
Trials        = Ones
Coefficient Estimates
Label           Estimate        Std. Error     Est/SE    p-value
Constant         1.80239         0.472703       3.813     0.0001
{F}SECTION[2]  -1.61000          0.500978      -3.214     0.0013
{F}SECTION[3]  -0.937523         0.541124      -1.733     0.0832
{F}SECTION[4]  -1.40044          0.499621      -2.803     0.0051
{F}NCCAT[2]     0.330888         0.383999       0.862     0.3889
{F}NCCAT[3]     0.940379         0.490486       1.917     0.0552

Scale factor:                  1.
Number of cases:             216
Degrees of freedom:          210
Pearson X2:              227.760
Deviance:                221.461
```

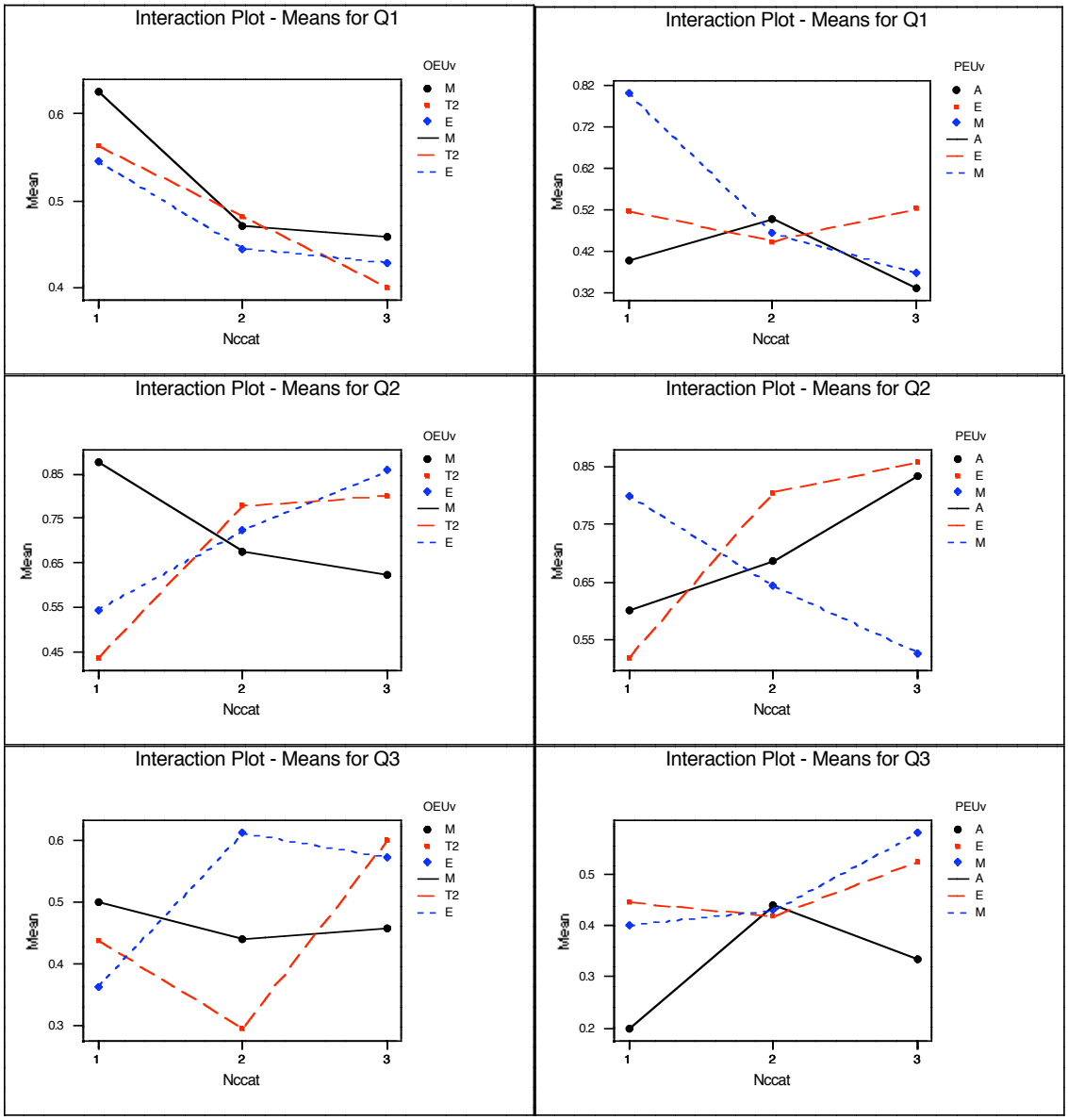
finished the semester in introduction to statistics. In this case, a student in section 2, was $e^{(-1.61)}$ = .2 times more likely to finish the semester than was a student in section 1.

According to the analysis, a HighNC student was, $e^{.94} = 2.56$, between 2 and 3 times more likely to be retained than a LowNC student. The p-values associated with the coefficients indicate that only the factors for sections 2 and 4 are significant. There is, however, the possibility that a follow-up study may find a relationship between NC categorization and retention in statistics.

3.2.2 Individual Multiple Choice Questions

The same type of analysis that was completed to examine the relationship between NC, EU and retention rate was applied to find differences in the percent of students correctly answering each of the multiple choice questions. The interactions plots, shown in composite in Figure 5.24, indicate the possibility of interactions between the constructs on the responses to the multiple-choice questions except possibly in the case of Question 1 and OEU. In that case, there appears to be main effects for NC, as NC rises, the percent of students answering correctly appears to fall, and OEU, with multiplists answering correctly at a higher rate than those in the T2 or Evaluativist categories. The largest interaction effect appears to be on Question 3 when NC and OEU are considered. The logistic regression analyses, however, reveal no significant main or interaction effects for differences in the correct response rate on any of the multiple-choice questions.

Figure 5.24: Exam Study Composite of Interactions Plots EU, NC and results on the Multiple Choice Questions



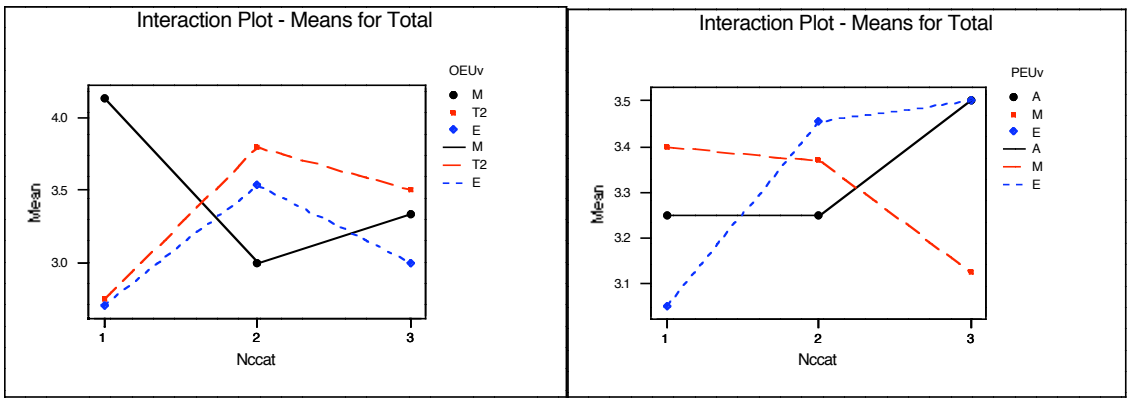
3.2.3 Total Score on Embedded Items

As discussed in Chapter 4, several long-answer questions were also embedded on the final exams. These questions asked the student to perform a hypothesis test and at

least one included a section asking students to discuss the reliability of the conclusion of the hypothesis test based on a stem plot of the data. Notes were made about the quality of the responses that students gave to these questions. Based on those notes, students were assigned a value from 0 to 3, where 0 points indicated the student could do little more than copy the information they had been given in the problem and 3 points indicated that the student could complete a hypothesis test without calculation and give a reasonable assessment of the validity of the conclusions based on the stem plot. One instructor did not include the stem plot item so the students in that section are excluded from the following analysis, leaving only 132 subjects.

The response of total points, from 0 to 6, the sum of those earned on the multiple choice questions and those earned on the long form hypothesis test questions, was no longer a binary response so this variable was analyzed differently from those described in the previous sections. Instead of a logistic analysis, the data was analyzed using both ANOVA and regression analysis techniques. Interactions plots were still used to infer the proper model for the data. They are shown in Figure 5.25 and indicate the possibility of a significant interactions factor, with both overall and physical world Multiplists behaving differently from subjects in the other EU categories as NC categorization varies. The

Figure 5.25: Exam Study Interactions Plots for NC, EU and Total Score



results of the ANOVA analyses did not indicate any significant factors or interactions. None of the p-values associated with any of the factors, other than section, was lower than 0.4 and many were nearly 1.

4. TARGETED STUDY

Recall that the purposes of the Targeted Study were 1) to assess differences in the development of strategic competence and statistical reasoning about hypothesis testing by beginning statistics students and 2) to find differences in the development of those aspects of proficiency that can be related to differences in Need for Cognition and Epistemological Understanding. There were ten Targeted Study subjects. All of these subjects submitted a writing sample for the Email from Dad task that was part of the Pilot Study. The next section is an analysis of the 40 writing samples. The subsequent section provides an analysis of the results of the 10 interviews.

4.1 Quantitative Analysis of the Writing Samples

The forty writing samples written in response to the Email from Dad by the Pilot Study and Targeted Study subjects were coded for the variables uncovered by the qualitative analysis discussed in the Pilot Study results. No new categories were generated by the Targeted Study responses. The eight dimensions along which the samples were coded were, 1) decisions about the medication, 2) definition of experimental design vocabulary, 3) assessment of experimental design elements, 4) discussion of sample size, 5) discussion of side effects, 6) discussion of anecdotes, 7) discussion of p-value, and 8) inclusion of personal statements unrelated to the medication. Axial coding generated subcategories for each category. The results of the axial coding were used to assign values to the subcategories so that the data could be analyzed using quantitative methods.

The following values were used in the coding:

Decisions about the medication (3 levels)

-1: Take grandma off the medication

0: No explicit decision about the medication

1: Leave grandma on the medication

Definition of experimental design vocabulary (3 levels)

0: No definition attempted

1: Definition of one element attempted

2: Definition of both elements attempted

Assessment of experimental design elements (3 levels)

0: Neither element is assessed

1: One element is assessed

2: Both elements are assessed

Discussion of Sample Size (3 levels)

-1: Sample is too small

0: Sample size not mentioned

1: Sample is of reasonable size

Discussion of Side effects (binary)

0: No mention of side effects

1: Side effects are mentioned

Discussion of Anecdotes (3 levels)

-1: Anecdotes are mentioned as worrisome

0: Anecdotes are not mentioned

1: Anecdotes are mentioned as not worrisome

Discussion of p-value (5 levels)

-2: Correct qualitative interpretation

-1: Incorrect qualitative interpretation

0: No mention of the p-value

1: Incorrect quantitative interpretation

2: Correct quantitative interpretation

Personal Statements (binary)

0: No personal statements made

1: At least one personal statement made

Classification of writing samples along the dimension of the discussion of p-values required me to make an assessment of what the subject had written. The ways in which p-values were described in the writing samples was discussed in detail in Section 1.4.4 of this chapter. Only one subject, who will be discussed in more detail below, gave a correct quantitative interpretation of the p-value in her writing samples. All other quantitative descriptions of the p-value were classified as incorrect. The reasonable and strong qualitative interpretations of the p-value listed in Section 1.4.4 of this chapter were classified as correct qualitative interpretations. Other qualitative interpretations were classified as incorrect.

4.1.1 Theory of Cluster Analysis

Discriminant analysis is one type of analysis that can be used to explore whether data predict subject groupings. In a discriminant analysis, the data are used to create a discriminator function. In this case, the function should sort subjects into either EU or NC categories, or into the interaction categories. Discriminant analysis, however, is not robust to deviations in the multivariate normality of the distribution of the data. Since the

coded variables will not have a multivariate normal distribution, such an analysis would not be appropriate for the data (Johnson and Wichern, 1998).

There is a statistical technique for grouping that may be used with data of any type and with any distribution, called cluster analysis. Cluster analysis is actually a collection of multivariate techniques used to categorize sets of data into previously unknown groups. The following discussion of cluster analysis is a summary of that presented in Kaplan (2001). The letter 'k', by convention, refers to the number of categories, or groups, formed by the clustering process. There are two main types of clustering algorithms, hierarchical and partitioning. Hierarchical algorithms are characterized by the fact that they consider every possible value of k, from one cluster containing all the data, to n clusters containing one datum each. A shortcoming of hierarchical methods is that, through the step-by-step process, once two elements have been joined (or split in divisive algorithms) they cannot be split (or rejoined).

Partitioning algorithms partition the data into k groups with the requirements that each group have at least one point and each point belong to exactly one group. An initial partition is specified or randomly generated and points are moved to different clusters if such a move reduces the overall 'within group differences' for the data. Every point is considered in turn, until a pass over each data point results in no movement of points between groups. In contrast to the hierarchical methods, group membership of every point is reconsidered until a quiescent state is reached. A shortcoming of partitioning algorithms is that the number of clusters is an input variable. In other words, it must be specified before the algorithm can be applied to the data. The subjects have been classified by EU and NC, so the number of groups is known. There is no difficulty, therefore, in specifying the number of groups into which the data should fall. In

particular, there were three groups for both EU and NC as main factors and eight groups when the interaction was considered. There were no LowNC-Transitionalists in the sample.

Once the choice to use a k-means partitioning algorithm has been made, one must still choose a distance metric and decide whether to standardize the data. Correlations between the variables suggest the use of the Mahalanobis distance metric (Everitt, 1993). Unfortunately, software that allows the use of Mahalanobis in a k-means clustering subroutine could not be found so all analyses were done twice, once with the Euclidean distance metric and once using taxicab distance. Crossing the two distance metrics with the two possible number of groups, 3 or 8, resulted in four possible cluster analyses.

The following rules for transforming data are generally used: Variables should be standardized if the measurement units are not well chosen and the investigator wishes to assign equal weight to each variable; Variables should not be standardized if the measurement scales are meaningful and the investigator wishes to retain the weights inherent in the raw data. Weights may be assigned based on prior assumptions about the data or knowledge of the domain. In this case, the coded variables are on roughly the same scale and have been well chosen by the researcher. Therefore, the variables were not standardized.

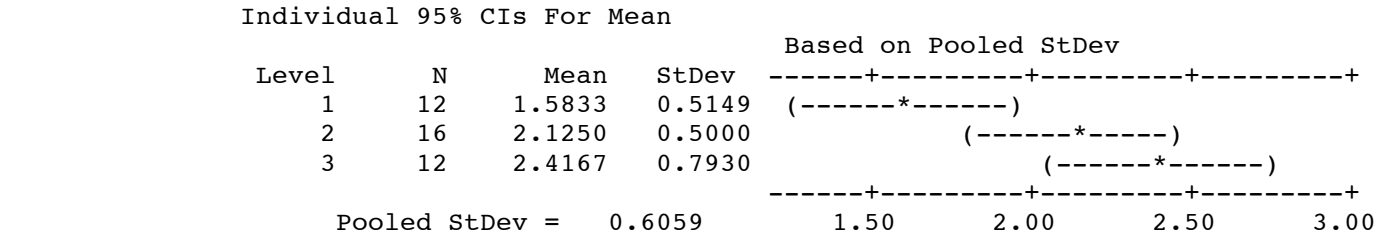
4.1.2 Cluster Analysis Results

None of the four cluster analyses was successful at partitioning the participants into either EU or NC groups of the category created by the interaction of the two groups. The analysis that came the closest to a reasonable grouping was the three-group solution using Euclidean distance. The results of this analysis begin to separate the subjects into EU categories. Seven of the twelve subjects placed in Group 3 were HighNC; the other 3

HighNC subjects were mis-categorized into Group 2. In addition to the 3 HighNC subjects, Group 2 contained 12 MidNC subjects and 1 LowNC subject. Eight MidNC subjects were mis-categorized into Group 1 and two into Group 3. Finally, five of the seven LowNC subjects were categorized into Group 1. One was mis-categorized in Group 2 and two into Group 3.

An ANOVA analysis of the differences in mean NC score by cluster found significant results (F = 5.88, p = .006). Figure 5.26 indicates that Group 1 has a significantly lower mean NC than does Group 3 with Group 2 in the middle. While the results do not constitute proof that differences in responses on the Email from Dad task can be attributed to NC, they suggest that perhaps future study with more subjects is warranted.

Figure 5.26: Targeted Study ANOVA of NC by Cluster



4.2 Qualitative Analysis of the Writing Samples and Interviews

Four major themes emerged from the qualitative analyses of the writing samples and interviews: 1) discussions of experimental design elements, 2) subjects' reactions to unbelievable conclusions, 3) types of evidence valued by the subjects, and 4) descriptions of p-values. A secondary discussion is embedded in the theme of experimental design. On the topic of Katherine Wallman's "mis-es," subjects both exhibited the tendencies and inferred that the general public would have a propensity to behave in the manner

anticipated by Dr. Wallman. Each of the four themes is discussed below along with a consideration of the implications for instruction suggested by the findings.

4.2.1 Experimental Design

Twenty-three of the forty subjects who provided writing samples to the Email from Dad task did not spontaneously explain what either "randomized" or "double blind" meant or why these factors might be important. Of the eleven subjects who attempted to explain the importance of the experimental design elements, only 1 made an attempt to discuss the value of having randomized the subjects. Many comments about the value of a blind or double blind experiment were limited to general statements like, "They do this so there won't be any bias in the experiment" (Pilot Study Subject 24). Only a few subjects made more specific statements such as, double blind "is needed to ensure that he (the doctor) doesn't see effects that he wants to see" (Pilot Study Subject 22).

In addition to the lack of attention paid by the subjects to the design elements, some of the writing samples show the development of some of the "mis-es" discussed by Katherine Wallman that were presented in Chapter 1. Recall that the "mis-es" are misunderstandings, misperceptions, misgivings, and, most importantly when considering the writing samples, mistrust. The most egregious case of mistrust was evidenced in the writing sample of Pilot Study Subject 14, a LowNC-Mulitplist with a 2.3 G.P.A. who expected to earn B in statistics. He wrote:

> Because the conducted experiment was a double blind experiment and what that means is that both the experimenters and subjects did not know what was given or taken. So the experiments did not know whether they gave the subjects the Makemewell or placebo, and the subjects did not know if they were taking the placebo or Makemewell. Even worst (sic) it could have been something else. … So until a real study is conducted for Makemewell, I believe that she should not be put on the treatment.

This subject is concerned that, in the use of a double blind experimental design, no one involved with the study will know what medications are being given and to whom. He accused the researchers of perhaps giving something that was not Makemewell to the participants. This pre-pharmacy student has developed mistrust for one of the strongest experimental designs used in drug testing today, based on a misunderstanding of the data collection procedure.

Pilot Study Subject 1, a MidNC-Transitionalist with a 3.65 G.P.A. who exhibited a high level of competence in procedural fluency and conceptual understanding, provides evidence that even people who do not mistrust the process may attribute some of the strength of a double blind design to the untrustworthiness of the researchers. He wrote, "the study conducted is very reliable since it chose people at random, meaning that everyone had an equal chance of being chosen and it was double blind, so even the scientists did not know who took the drug so their results could not be made up." The end of this statement implies that the subject had some notion that researchers may fabricate or manipulate their data in order to receive the results they desire.

The responses by the Targeted Study subjects relating to experimental design were, in general, stronger than those that had been given by the Pilot Study subjects. It should be recalled, however, that the Targeted Study subjects represented a class of students who had demonstrated conceptual understanding and procedural fluency in statistics. The Targeted Study subject group was far more homogeneous, based on both G.P.A. and grade in statistics, than was the Pilot Study subject group. This may account for some of the differences that will be discussed.

Only Kyle and Natalie, of the Targeted Study group, were found to be weak in their ability to discuss experimental design. Kyle could define both "randomized" and

"double blind," but, when asked why these features were important, he said only that they will keep patients from thinking that the drug is working. He did not explore problems or issues of experimental design when he questioned study results, even when probed to do so. Natalie was unable to even define "randomized" and "double-blind." Most of her interview was spent with her asking questions about procedures that were used in the studies; she never assessed their value.

The other eight Targeted Study subjects were able to define and assess the experimental design elements in the Email from Dad task. In addition, they were able to assess the value of the inclusion of design features. They knew that choosing a random sample or randomizing the participants into groups controls bias and provides results that are more generalizable to the population. The subjects also considered experimental design factors when they did not believe the outcome of a study. This happened frequently in discussions of the ESP Study. While the subjects tended to be inexpert in creating arguments based on design elements, they did seem to know that experimental design is a source of bias and a place to look when one wants to critique a study. Subjects were less likely to discuss experimental design spontaneously in conversations about the Lyon Diet Study, the results of which were less controversial in general.

Returning to the discussion of "mis-es" and experimental design, there is evidence in the Targeted Study sample of mistrust of researchers. Of the eight subjects who talked about a double blind design reducing bias from the researchers, only three were clear that the bias could occur without intent. These subjects used the words, accidental and unconscious, to describe the actions of the researchers. The strongest mistrust of researchers is exhibited in Mariposa's interview. She said:

> Well, if you are a researcher and you want your medication to get off, you are going to sit there and try to tweak things and make it look good. But with what

> they did they, they are making it as honest as possible and they are not trying to lie through their evidence and make it not truth.

Megan made a similarly strong statement; the other subjects were less emphatic or clear about the relative intent on the part of the researchers.

Sarah's transcript provides an interesting comment on "mis-es" in her discussion of the relative value of the results of the Lyon Diet Study when compared to those of diet articles she reads online. She said, "I like this (the Lyon Diet Study) more but some other people might see it as, 'Oh, they are giving me all these numbers and lingo and they are trying to trick me.'" She went on to say that one person's account of a diet that worked for her might seem to most people more believable than numbers. Sarah concluded by saying:

> I personally like statistics more but as we learned in class there is a lot of ways to twist the statistics to make it look like something that it's not. So I like statistics more but I know there is(sic) also some instances where you have to watch out.

In her discussion, Sarah seemed to agree with Dr. Wallman's claim that the general public tends to distrust statistical results. She also appeared to have developed a sense of the need to use critical thinking when assessing statistical results.

4.2.2 Reactions to Unbelievable Conclusions

Nearly every subject in the Targeted Study found the results of the ESP Study to be unbelievable. The subjects who did not believe in ESP were surprised by the results for the dynamic pictures and the subjects who believed in ESP were surprised by the results for the static pictures. Two subjects, Kefira and Hannah, were also shocked by the number of people in the Lyon Diet Study who died during the study. For these subjects, the context of the task created different responses than did tasks with contexts that led to believable conclusions. There were three main themes to the reactions subjects had when

faced with an unbelievable conclusion: 1) critique the design of the experiment, 2) request more information, and 3) consider a rational explanation for the study.

There was a general propensity of subjects to look for problems in experimental design when the conclusions were unbelievable. This response to unbelievable conclusions was unexpected given that few subjects spontaneously discussed experimental design in the Email from Dad task, which had believable study results.

Hannah and Charlotte, both of whom believed in ESP, also believed that ESP could not be studied scientifically. They both indicated that the use of novices as subjects might have diluted the study results. According to Hannah and Charlotte's reasoning, novice receivers would have been guessing randomly. If there were enough novice receivers, their responses, which would have been guesses not based on true reception, might account for the non-significant results in the case of static targets.

The main critique of the study given by those who did not believe in ESP was that there were only four targets from which the receivers chose. Megan suggested that there should have been 30 targets. Under the condition of 30 targets, subjects guessing randomly would be expected to be correct on roughly 3% of the trials. The minimum number of trials to meet the conditions for inference with 30 targets would be 450. In a study with 450 trials and 30 targets, the same p-value would result if 6% (28) of the trials were successful. The subjects who said that a higher percentage of successes would be necessary to convince them of the existence of ESP suggested that convincing percentages that ranged from 45 to 95%. This suggests that it is unlikely that significant results from a study with 30 targets would be perceived as having more value in general. Sarah suggests that some of the films might have been more eye catching and that would

have biased the results. Both Kyle and Bradford specifically mentioned that they could not find any flaws in the study.

The second type of response given by subjects in reaction to surprising results was a request for further information. The most common request for more information was to query about replication studies. In addition to that, Kyle, Natalie and Bradford asked clarification questions about the reported study. For example, regarding the ESP Study, Kyle asked how the researchers defined success and Natalie asked whether the senders and receivers knew each other. One unusual request came from Bradford who said that he would like to read critiques of the study. It is not clear whether he would have asked for this material if the results had been believable, but he seemed to have a very clear notion of the importance of counter-arguments in evaluating the evidence presented.

The third reaction to the surprising results was to search for a rational explanation for them. Bradford expressed concern about the ESP Study results because he could not think of a rational explanation for them. Four of the other subjects posited explanations that explain or negate the surprising result. Jewel said that she had heard that humans only use 18% of their brain so maybe ESP happens when a person can use more. Hannah also referenced brain function, suggesting that the difference in the results between static and dynamic targets might be the result of differences in the areas of the brain that process the two types of stimuli. Sarah noted that the static pictures should have been easier to choose correctly because they represent only one thing to be thinking of, while the movies are many images on which to be concentrating. Sarah also said that the results could just represent a good guessing day. Kefira's explanation was about her shock at the number of Lyon Diet Study subjects who died during the study. She said

that perhaps the subjects were relatively old. In that case, the deaths would represent a more reasonable number.

4.2.3 Types of Evidence

Five of the ten Targeted Study participants specifically mentioned in the interviews that statistical data is more valuable to decision making than are anecdotes. For example, Natalie wrote to her father that, "the doctor's words and statistic are more valuable and accurate than the neighbor's personal statement." Mariposa, in her discussion of the Lyon Diet Study, said:

> If I was presented with someone saying, on a Good Morning America show, "You should eat this instead of this." or I was shown this article where it's got the proof and they did an experiment, I would be much more willing to believe the article than some person just saying you should eat this and not this because it worked for me. I'm like, "what evidence is there for that? You lost some weight, what are you not telling?"

The interview subjects all specifically made mention of the value of study results as convincing evidence. In general, the subjects seemed to be able to differentiate between the value of evidence and anecdote. While this study cannot be used to imply that this ability to differentiate is a result of taking a statistics course, it does demonstrate that there are undergraduate students who have that ability.

In addition to study results, there were two types of arguments that subjects gave to support their ideas of a healthy diet: 1) preponderance of the evidence and 2) justification. Some subjects had learned what constitutes a healthy diet from their families or in school. Others talked about having heard of the benefits of olive oil, increased consumption of fruits and vegetables, or reduced consumption of red meat from other sources. Finally, there were subjects who talked about aspects of healthy diets that have "stood the test of time." Justification arguments to support healthy diet ideas were

generally based in biology. For example, Kyle discussed learning in genetics class that the blackened parts of meats have nitrates, which are detrimental because they may cause cancer in the DNA. Charlotte made an unusual justification argument about diets, explaining that in her nutrition class they analyzed the advertising for a diet to make conclusions about the possible effectiveness of the diet. If the claims made in the advertisement were unreasonable, for example, losing "10 pounds in one week," that would be evidence, for Charlotte, that the diet was not a viable weight loss option.

The subjects' notions about what makes a compelling argument were quite different when they discussed the ESP Study conclusions, which they find to be not believable. Neither Kefira nor Megan was convinced of the existence of ESP, but both were more open to the possibility based on the study results. None of the other subjects were convinced by the study results. Natalie and Sarah said they would be more convinced if the success rate were higher. For Natalie a convincing success rate would be roughly 50%; for Sarah the rate would have to be about 75%. Hannah, Jewel, Mariposa, Charlotte, and Kyle said that they would need to witness someone using ESP in order to be convinced of its existence. For example, Charlotte says that she would be convinced if she "saw it in real life, like, somebody doing it on a stage or something." She continued, "If I saw it with my own eyes, then maybe that would give me more evidence for it."

Jewel and Kyle both exhibited meta-cognitive behavior in their reactions to the ESP Study. Jewel, after saying that she would need to experience ESP to believe in it, immediately admitted, without prompting, that the experience "would be even less valid than any sort of psychological experiment." She then made an argument based on an authority saying that "Psych[ology] Bulletin is a good journal so that's another thing to

their credit. It's been checked through very diligently. So, that would give them more weight, but I'd probably always be skeptical." At the end of this section of the interview, Jewel created a rational justification for ESP, saying that she has heard that people only use 18% of their brain capability so maybe there are people who can use more. The people who can access more of their brain might have ESP.

Kyle was even more concerned about the hypocritical nature of his response to the ESP article. Early in the conversation about the article he said:

> It's just that it's something I don't really believe in absolutely. I'm a skeptic. But then again, I'm not sure. Well, this can go against me two ways. One way I can be like, "alright, so I have to accept that there is such a thing as ESP," which I don't, really, unless I actually see someone, who would be like reading my mind immediately. Or I have to say statistics is wrong which would be bad for the two things [tasks] I just did. But, I know statistics can't, is not wrong.

He returned to the theme of his stubbornness about refusing to believe the statistics in the case of ESP while completely believing them when they refer to drug trials several times during the interview. In the end, he maintained that he must observe ESP happening in order to believe in it. He reconciled this with the belief in statistics derived from pharmaceutical trials by saying that he has taken medications and received relief, which constitutes having an experience with medicine.

4.2.4 P-values

A quantitative response about a p-value was defined above as one in which the subject attempted to ascribe meaning to the value. Only 1 of the 16 subjects who attempted to explain, quantitatively, what the p-value of .001 meant about the medication and/or the study was able to do so correctly. Megan wrote:

> The way it works is they look to see what the chances are of getting the result of lower blood pressure (and the experiment did show) if in actuality the medication didn't work. This is the weird p-value (< .001) you sent me. If the Makemewell

> drug didn't lower blood pressure the result that the experimental group had would happen only .1% of the time (very rare indeed!).

By contrast, about half of the 13 subjects who explained the p-value to their father in writing were able to do a reasonable job of it. An example is given by Pilot Study Subject 28 who wrote, "The $p < .001$ means that the Makemewell is probably more helpful rather than not having a medication at all."

The interviews provide a richer fabric of what students are thinking about p-values and show that the situation is more complicated than the writing samples might suggest. Some of the Targeted Study subjects were consistent in their interpretation of the p-values and provide no interesting insight into the development of understanding of p-values. Two of the subjects, Megan and Bradford, had the same statistics instructor and were the only subjects in the study who had that instructor. Megan gave a consistently correct interpretation. Bradford claimed throughout his interview that the p-value represents the probability that the null hypothesis is true. Even with this persistent misconception, Bradford was the only Targeted Study subject to note that the p-value of .13 associated with the test of differences in side effects between the control and experimental groups in the Email from Dad task is perhaps not enough evidence of a difference. He explained that the value .13 means that there is "a 13% likelihood that the two groups could have appeared equivalent by chance." He continued, saying that the test is "not conclusive either way," and that "the drug could have side effects."

The way that Sarah and Kyle talked about p-values were similar to each other; they also had the same statistics instructor. Sarah specifically linked her understanding of p-values to the classroom experience. Sarah claimed that, while they always calculated the p-value in class, they used it to determine the "strength of something." She went on to say that after the p-value was calculated, if p was large, it tells them that "the hypothesis

is probably right" and if the p-value is low, "there is something not quite right about with our hypothesis and it's got to be something else." Kyle, who did not discuss the class explicitly, also said that the researchers must make an assumption at the beginning of the research. He stated that the p-value is the probability the assumption the researchers made was wrong.

The similarity in the discussions from two students cannot be extrapolated to other students in the class. Hannah, who had the same instructor, remembered that the instructor always drew a "bell curve chart" in which the tail areas are the "unlikely results for happening" and the body area represents the way a "majority of…people respond to things." She appeared to believe that the "chart," which represents the distribution of all possible sample means, actually represents either the individuals in the sample or in the population. Her comments on p-values beyond this discussion were limited to labeling those p-values under 0.1 as "significant" and those above 0.1 as "not significant." Charlotte, who is the only other student of the same instructor, used the labels "significant" and "not significant" in the same way that Hannah did. When asked directly what the .05 meant as a p-value in the Lyon Diet Study, her response was similar to those of Sarah and Kyle. She said that there is a 5% chance that you should reject the hypothesis that the diet helps with cancer because it is wrong.

Sarah's description of the class implied that her instructor taught inference using significance levels. Charlotte gave more evidence for this claim when she asked, in her interview, whether the doctor gave an alpha level for his test of side effects. She said, "when you say, significant or not significant, you have to give an alpha level." Student responses on the final exams further confirm this claim. If calculated p-values were always compared to a specified alpha level, these students may not have discussed the

quantitative meaning of the p-value. The responses given by these students illuminate conceptual misunderstandings about the meaning of the p-value that may result when inference is taught from the significance level viewpoint. Furthermore, a comment made by Sarah indicates another "mis" that may develop when this treatment of inference is the predominant one used in a statistics class. She said that a researcher could pick any level of significance and then say the results are not significant, but they can still have practical significance.

Kefira provides an example of a student who was not taught inference solely from a significance level viewpoint. Her first comment about p-values was that low p-values provide support for the null hypothesis. While this is an incorrect statement, it shows a different perspective about the p-value than was provided by the students described above. When asked what a p-value is, Kefira said, "you get a test statistic, and then a p-value, and then a conclusion. So the p-value is basically telling…you how significant your results are." Later she said, "After you get a p-value of .1, … you can't really make a confident statement about anything." Kefira said about the results of the ESP Study, "there is some evidence that static and dynamic pictures don't produce the same ESP results." Based on having worked with the instructor of Kefira's class, I believe that Kefira demonstrated procedural fluency as it would have been tested in her course. The conclusion she provided is exactly the conclusion that would have been expected by her instructor on an exam or homework and the way she discusses the process and the levels of significance closely matches my experiences observing Kefira's instructor.

In addition to the "expected" responses Kefira said, "a really, really low p-value means that one thing you are testing for is true," and a high p-value means "the alternative of what you thought was true." This resembles the claim made by some of the

students who were taught using a significance level approach, that is the p-value gives the probability that the hypothesis being tested is true. Kefira, however, ascribed meaning to the value differently. For Kefira, the p-value represents the percent of the time that what they thought would happen actually happened. This is probably a misconception arising from either a misunderstanding in relating the sample distribution to inference or confusion between hypothesis testing and confidence intervals.

Mariposa provides evidence for what students may be thinking when they use the phrase "by chance" in discussions of p-values. She both wrote and said that a small p-value means that the "results were outside of chance." Later Mariposa said that the small p-value found for dynamic pictures does not make sense to her because "there is so little chance involved in getting that result." This statement and the sentiment behind it appear to be close to the true meaning of the p-value. Furthermore, when asked to incorporate the p-value .03 from the Lyon Diet Study into a sentence about diets, Mariposa said:

> There's only a 3% chance that your regular diet…that there'd be no effect on your [results], that means that there is a 97% chance that this [diet] is good stuff. Well, not a 97% chance, it's just that it's more likely that it's not random chance that is doing it but that the actual experiment is what is making the difference in these people.

At the end of her statement Mariposa makes an essentially correct assumption. A small p-value tells you that it is unlikely that the results were due to chance and, therefore, more likely that they were due to the condition that was imposed.

After I realized that Mariposa's written statement using the phrase "by chance" was the foundation of an essentially correct understanding of the p-value, I reread the writing samples to reanalyze the response of the subjects who used the phrase "by chance" in their writing. The interpretations of the p-values of these subjects had been classified as quantitative and incorrect. Mariposa's response prompted me to consider

whether the responses of those subjects illuminated a place that students stop on the way to developing a deep understanding of the p-value. There were five such subjects. On re-reading the transcripts, I found that the part of a correct interpretation of a p-value that was missing from these responses was the acknowledgement that the probability is conditional on the null hypothesis being true. For example, Subject 26 wrote that the p-value meant that the doctor's "study results would happen 1 time out of 1000 by chance." If the subject had written that the p-value was meant that the doctor's would happen 1 time out of 1000 by chance if Makemewell were not more effective than the placebo, it would have been considered to be a correct interpretation. Without more information from this Subject, it is not possible to ascertain whether she knows that the probability is conditional. This reiterates the value of interviews in educational research and of alternative assessments in the classroom to find out what students know and believe.

Chapter 6: Reflections

This research set out to provide an initial dispositional attribution model to describe differences in the development of statistical proficiency. The dispositions that were chosen to be the basis of this model were Need for Cognition (NC) and Epistemological Understanding (EU). The quantitative analyses did not find a significant relationship between EU, NC and conceptual understanding and procedural fluency of hypothesis testing as assessed by the Exam Study questions. This was not an unexpected finding. What was unexpected was the lack of findings relating EU and NC to differences in the development of statistical reasoning and strategic competence. While the choice of the constructs was well supported by the literature, the results of the study may be limited based on that choice. Furthermore, while the experimental design was expected to net a large subject pool for the Targeted Study, this was not the case. The research did identify several common themes in the student discussions of hypothesis testing. The first section of this chapter discusses the findings related to the common themes and their implications for the classroom. The second section of the chapter provides a brief discussion of the limitations and delimitations of this work. The chapter concludes with a discussion of future research suggested by this work.

1. Findings and their Implications For Teaching

Three findings of the research are discussed below. The first three findings pertain to, 1) how beginning statistics students discuss experimental design issues, 2) what they consider as an evidentiary basis for claims, and 3) what they understand about p-values.

1.1 Experimental Design

A list of items that would assess the statistical reasoning facet of statistical proficiency was given in Chapter 4. One of the items mentioned on that list was: students should be able to discuss strengths and weaknesses of an experiment and comment on how the experimental design impacts the strength of conclusions that can be made. The results presented in Section 4.2 of Chapter 5 suggest mastery of this skill by beginning statistics students may be weak. In particular, they tend spontaneously discuss experimental design features only when the conclusion presented is unexpected or counter to a previously held opinion. Further, when students are probed to discuss features of experimental design, they tend to exhibit a mistrust of researchers in general.

Alacaci (2004), through interviews with statistical experts, concluded that knowledge of experimental design "constitutes the backbone" of an expert knowledge of statistical techniques. Based on his findings, Alacaci suggests that the teaching of inferential statistics should include explicitly a link to experimental design in order for novices to develop a knowledge base from which they will be better able to become experts (Alacaci, 2004). The difficulties exhibited by the subjects with regard to experimental design are, therefore, disappointing.

While the findings with regard to student discussions of experimental design are disappointing, they are not unexpected. Belief Bias, as discussed in Chapter 2, predicts that people will tend to rate a logically valid argument as less strong than a logically invalid argument when the logically valid argument leads to an unbelievable conclusion and the logically invalid argument leads to a believable conclusion. In an empirical study of graduate students in science, Koehler (1993) found that his subjects tended to give more favorable ratings to research reports that made conclusions with which they were in

agreement (called "belief-congruent" findings). The differences between the judgments were not based on harsh critiques of "belief-incongruent" results. This led Koehler to conclude that scientists are more likely to assume that a study with "belief-congruent" findings has been correctly conducted, but when study results are "belief-incongruent," scientists' skepticism is activated. A follow-up study of practicing research psychologists and scientists provided "further evidence for an agreement effect in…judgments of evidence quality" (Koehler, 1993, pp. 46) by experts.

The findings with regard to experimental design, while perhaps not surprising, present the following challenge for statistics educators: How can we encourage students to act with the same skepticism about experimental design whether the results of the study give evidence that either confirms or is in conflict with the opinion held by a student without instilling in them the sense that all statistical results are based on untruths or explicit attempts to have skewed the results? Alacaci (2004) gives four suggestions for improving the teaching of experimental design as part of the curriculum on inferential statistics: 1) make associations between statistical techniques and design aspects; 2) use think aloud techniques while completing examples of inference that include experimental design issues in the out-loud-thinking; 3) use examples of actual research in teaching inference; and 4) construct scenarios for teaching examples in which there are experimental design issues that will provide valuable class discussion.

The relative lack of understanding of experimental design shown by the Pilot Study subjects may be partially attributed to the treatment of experimental design by the textbook the subjects used in their statistics course. Experimental design is covered in two chapters of the textbook. The problems in the textbook designed to assess understanding of experimental design are largely exercises in using the random digits

table to select a sample or in drawing an outline of the experimental design. These two chapters tend to be covered around the time of the first exam in the course. Students spend the first few weeks learning basic exploratory data analysis techniques, finding summary statistics and graphing techniques for one and two variable data and learning about the normal distribution. After the experimental design material, students begin a brief study of probability leading to material on sampling distributions.

The chapters on inference, which might include the strength of conclusions that may be made based on the design features of a study or experiment, are not covered until after the second exam. In the discussions of conditions for inference, attention is paid to the selection of a simple random sample, the size of the sample and conditions about the distribution of the sample. For example, that the distribution of the sample data should not appear too skewed for the sample size or have outliers. Ideas about design are not incorporated into the discussions about conclusions that can be made based on inference. An example of a textbook that does incorporate discussion of experimental design in the sections on inference is The Statistical Sleuth: A Course in Methods of Data Analysis by Ramsey and Schafer (Ramsey and Schafer, 2002). Based on the findings of this study and the suggestions made by Alacaci, it is suggested that students may benefit from the inclusion of specific attention to the connections between experimental design factors and the scope of a conclusion that can be made from and experiment in the curriculum for the introduction to statistics course.

1.2 Evidentiary Basis

The Targeted Study subjects mentioned three types of evidence as being convincing: 1) statistical results, 2) a preponderance of evidence, and 3) a justification or rationalization. When the conclusion suggested by statistical evidence was incongruent

to a prior belief, subjects tended to be less convinced by the statistics. In this case, subjects tended to search for a justification for the conclusion or rely on their pre-existing opinion. These opinions tended to be based on a preponderance of evidence. It should be noted that reliance on prior evidence is a reasonable reaction from a Bayesian perspective if the prior knowledge has an evidentiary basis.

Current research in psychology on people's criteria for justifying causal claims predicts that people tend to value theoretical explanations over evidence (Kuhn, 2001). Kuhn (2001) states that the literature suggests that people "depend on explanations that allow their claims to make sense to themselves and others" (pg. 1). The tendency of the Targeted Study subjects to want an explanation for the unexpected outcomes is consistent with the Kuhn's findings. Kuhn further suggests that recent research indicates that the "preference for explanation over evidence is dependent on context and on the strength of the evidence" (ibid) and that it diminishes developmentally, disappearing in among highly capable undergraduate students. The data provided by the Targeted Study interviews suggests that even high functioning undergraduates rely on explanation over evidence in the justification of claims when the claims are counter to a previous belief.

The Cobb Report (1992) specified as a goal for introductory statistics classes the development of statistical thinking and mentions, in particular, "recognizing the need to base personal decisions on evidence (data), and the dangers inherent in action on assumptions not supported by evidence" (GAISE, pg. 3). The comments made by the Targeted Study subjects when discussing unexpected findings indicate that these students have not met the goal of recognizing the value of evidence in this case. The challenge for statistics educators, therefore, is to help students develop an appreciation for statistical evidence even when it does not support a pre-existing belief.

This finding suggests some specific suggestions for instruction. If all classroom examples and homework problems in a statistics course are about contexts with which students are unfamiliar or lead to unexpected results, students may not accept the statistical process and reasoning as valid. Statistics instructors should consider the contexts of the examples and problems they use in class. Students should be familiar with the contexts of the problems. In addition, instruction should include some examples where the conclusion is entirely believable. Instructors should actively engage students in discussions of both believability of conclusions and the types of arguments they find convincing.

1.3 P-values

The first finding about student understanding of p-values that can be made from the data collected for this work is that context matters. More specifically, the results of the problems in the Pilot Study that asked participants to evaluate interpretations of a p-value as valid or invalid showed that students are more likely to correctly identify a valid interpretation of p-value when it is presented without context. In fact, no subjects identified the correct interpretation of a p-value in context without having identified it correctly without a context. The finding that context affects how students answer questions about p-values extends the findings about how students discuss experimental design and the types of evidence they find convincing. It adds to the basis of the suggestion that instructors should be mindful when choosing contexts for classroom instruction.

The next finding about student discussions of p-value is that students in the same course tend to develop many different definitions and understandings of the p-value. This

result may seem surprising given the relative lack of variability found in other aspects of the development of statistical proficiency. In fact, the variety of misinterpretations of the concept of "p-value" is supported by the literature. Empirical studies in the fields of education, psychology, statistics and mathematics and statistics education have documented at least thirteen different misconceptions of the p-value concept (Lane-Getaz, 2005).

The number of ideas about p-values that are developed spontaneously by students presents a challenge for statistics educators. Given the number of misunderstanding that can occur, how can we help all students to move in a direction that will provide them a better understanding of p-value? It is beyond the scope of this work to answer that question. Some of the other findings from this research, however, may provide direction for future research to investigate the question. The writing samples indicated that perhaps students were more adept at discussing p-values qualitatively rather than quantitatively. To put it another way, it appeared as though students had the general idea of what a p-value means and what one can conclude from it even when they could not give the correct quantitative interpretation. This initial finding suggests that perhaps there is value in helping statistics students develop a general sense of the meaning of the p-value prior to attaching quantitative significance to it. Based on this I would have suggested that significance level testing should be made more salient in the curriculum when inference is introduced because it would allow students to develop a more conceptual understanding of the p-value before attaching a probabilistic meaning to it.

The students who were taught hypothesis testing from a significance level viewpoint, however, had difficulty in explaining the meaning of the p-value. Furthermore, they tended to exhibit mistrust of statistical research based on the arbitrary

nature of the choice of an alpha level. These results suggest that an instructor should be cautious in the use of a significance level approach to the teaching of hypothesis testing. More generally, these results, paired with differences seen among other instructors' student groups suggest that the role of the instructor is important in the development of student understanding of the p-value. It may be the case that instructional approach, rather than instructor, is the mitigating factor of the observed differences. Since the classes from which the subjects were taken were not observed, this research cannot distinguish between the two factors. This distinction could, however, be studied in future research.

This research delineates different ideas held by students about p-values but cannot be used to make conclusions about the types of understanding that develop as a result of instructional decisions. Therefore, these findings raise questions that cannot be answered by this reseach. Some specific questions that that are derived from this research are: What type of instruction or assessment helped Megan to be able to state the precise quantitative meaning of the p-value? and What type of instruction or assessment helped Bradford develop the ability to think critically about the p-value?

The last finding of this research about student understanding of p-values is that students who state the meaning of the p-value without specifically mentioning that the probability is conditional on the truth of the null hypothesis may, in fact, know that the probability is conditional. Students might neglect to include the statement of the conditional because they feel it is "understood," or they may truly not understand that the p-value is a conditional probability. This has implications for teaching and assessment. Instructors should use instructional techniques and assessments that are designed to address the conditionality of the p-value.

The writing samples in this study did not discriminate between correct and incorrect understanding of the conditionality of the p-value. This suggests that standard written assessments may not be sufficient to discriminate student understanding. This idea is not a new one in statistics education. Garfield and Gal (1998) discuss the advantages of alternate forms of assessment in statistics. The authors suggest that the use of assessments such as, projects, portfolios, concept maps, and critiques of issues in the news, provide "more complete information about what students have learned and can do with they knowledge" (pg. 4) and will provide a "richer and more complete representation of student learning" (ibid).

2. LIMITATIONS AND DELIMITATIONS

2.1 Framework for Statistical Proficiency

"All models are wrong, but some are useful" George Box

The model that describes statistical proficiency, presented in Chapter 2 of this work, has been useful for the purpose of this research. A limitation of the framework is that it has not been validated empirically. It is not known how closely the model matches the reality of either student learning or expert knowledge. Using Box's terminology, we do not yet know how useful it is in practice. The process for validating the framework has begun. I have solicited feedback from statisticians and statistics instructors in an initial attempt to find out whether the model accurately describes what they have experienced and to ascertain whether the aspects of proficiency have received correct emphasis. The full process of validating the framework is an aspect of future directions for this work and will be described in more detail below.

2.2 Test-Retest Reliability of EU and NC

Eight of the ten Targeted Study subjects changed classification on one or both of the dimensions, EU or NC. NC is positively correlated with educational level (Cacioppo, et, al., 1996) and EU is theorized to be a developmental construct (Kuhn, et al., 2000). The subjects were college students, and differences in retest categorization may have been developmental in nature. They should, therefore, not have been unexpected. Of the seven subjects who were re-categorized on the EU scale, however, three were in the regressive direction. Two of the three subjects who changed NC groups regressed to a lower categorization. The possibility of retest reliability problems with the NC and EU tasks creates a limitation for this study.

In order to find out whether the tasks that assess EU and NC might have re-test reliability issues, two keyword searches on the PyscINFO database were performed. The first was a search for "epistemological understanding" and "reliability" and the second was "need for cognition" and "reliability." There were no articles about EU that discussed reliability and 11 that discussed reliability of NC. When the search was modified to include "retest" instead of reliability, there were only five NC articles that investigated retest reliability, only one of which was completed using the English version.

The one report of re-test reliability of the English version of the 18-question short from for NC is a four-paragraph article. It states that when 71 undergraduate students in a lower level psychology course were tested for NC on two occasions that were seven weeks apart, the correlation between the scores was .88 (Sadowski and Gulgoz, 1992). The correlation coefficient, r, is a measure of the strength of the linear relationship between two variables. It is not a measure of reliability. If, as Sadowski and Gulgoz claim, NC scores were relatively stable, we would expect the differences between scores

to be small and the distribution of these scores to be roughly symmetric and centered at zero. This assumption would be better tested through the use of a matched pairs two-sample t-test for means.

A better use of regression analysis to measure the retest reliability of the NC task might be to use the following procedure. If the task has retest reliability, then we would expect the scores on the two administrations to be roughly the same for each subject. The least squares regression line (LSRL) would then be roughly the line $y = x$. The correlation coefficient does not give us any information about the coefficients of the LSRL. Sadowski and Gulgoz (1992), however, provide enough information in their paper for us to calculate the LSRL:

First administration: $m = 23.41$, $SD = 22.76$

Second administration: $m = 20.83$, $SD = 24.62$

Therefore, the LSRL, where y represents the NC score on the second administration of the NC task and x represents the NC score on the first administration of the NC task, is: $y = .95x - 1.40$. While this result is nearly the hypothesized result of $y = x$, caution should be used in interpreting the results. Without access to the raw data, it is impossible to check whether the conditions for inference are met. In light of the above discussion, the re-test reliability should be studied in more detail prior to using EU and NC in future studies as measures of dispositions. Possible methodologies for doing so will be discussed in the section below on future directions.

2.3 Experimental Design and Enactment

After the population study had been completed, there were a total of 216 subjects who had agreed that they could be contacted to participate in future studies. This appeared to be quite a large subject pool from which to draw for the Targeted Study.

However, the target groups for that study were the High and Low NC, Evaluativists and Multiplists. Among the 216 possible subjects, there were 8 HighNC-Evaluativists, 17 LowNC-Evaluativists, 30 HighNC-Multiplists and 14 LowNC-Multiplists. The subject pool was limited further after students left the course or did not demonstrate procedural fluency and conceptual understanding of hypothesis testing. When a possible subject list was created in order to contact potential subjects, it contained 4 HighNC-Evaluativists, 3 LowNC-Evaluativists, 6 HighNC-Multiplists and 3 LowNC-Multiplists. When it was suspected that either the assessment tasks were not reliable under retest conditions or that subjects had changed categorization over the course of the semester, the High and Low NC, Transitionalists were invited to participate; there were 9 HighNC-Transitionalists and 4 LowNC-Transitionalists. The 10 subjects who were interviewed represent more than one-third of the possible subjects, but not nearly the number that it had been intended to study.

The design of the study was intend to reduce sources of variation. By studying only students who had successfully mastered procedural fluency and conceptual understanding, the research would focus only on differences in other aspects of statistical proficiency. Unfortunately, this design limited the size of the sample. This is a delimitation of the study because it limits the generalizability of the conclusions. In this case, the findings from the interviews may not represent all of the types of understandings and misconceptions that are developed by beginning statistics students, even those students who are like the students in the sample. In order to solve this problem in the future, studies could begin with a larger population from which to draw subjects.

Another delimitation of the study is the relative homogeneity of the subjects. All are relatively good students who did well in their first statistics course. This may be the

cause of the relative lack of variation in the interview responses. It would be interesting, in the future, to interview students with different backgrounds about the same tasks. This would take the form of a cross-sectional study, in which students across all levels of statistics course taking, for example undergraduate statistics majors and graduate students in statistics as well as non-mathematics majors, would participate. Differences and comparisons in understandings and misconceptions in statistics could then be done across levels.

A final delimitation of this study is that it is not known how expert statisticians would react to the interview tasks. The responses given by the students cannot be compared to expert statements. The results of this study would be strengthened by a study of these tasks using statisticians. The statisticians could either be interviewed in the same way that the students were or they could be asked to provide responses they would expect from a good student if they had given a class of students an assignment to read the article and provide a written critique of it. This would provide a better measure against which to compare the responses of the Targeted Study subjects.

3. FUTURE DIRECTIONS

3.1 Framework to Describe Statistical Proficiency

The first step toward the validation of the framework would be for statistics instructors and statistics educators to discuss and critique the model in order to reach consensus in the field about aspects such as, the inclusion and emphasis of elements of the model and vocabulary and definitions (Murray, et al., 2005). This process has already begun. The brief description of the model presented in Chapter 2 has been presented in talks and sent to several statisticians and statistics educators for comments. As of this

writing, no feedback has yet been received. I plan to present this model at an upcoming national statistics education conference to further this step of the validation process. At that point, the field would have a "common language and a vehicle for discussing the definition of the skill area" (Murray, et al., 2005).

The next steps in validating a framework are to identify task characteristics for each aspect of the framework and then to decide how those characteristics can be used to construct tasks (Murray et al., 2005). Once tasks have been constructed and an instrument to assess statistical proficiency has been developed, this instrument will also need to be validated. There are several parts to the instrument validation process. The instrument should have internal reliability, such as that discussed in Chapter 5 (Nunnally, 1967). Furthermore, a principal components analysis should indicate support for the theory behind construction of the assessment instrument. It should find the number of components to be equal to the number of aspects of proficiency that are being tested. The component of each proficiency aspect should be a weighted average of the responses to the tasks that, in theory, are being examined. Any components that appear to be contrasts between two competencies may indicate interactions in the development of statistical proficiency.

There is already interest in the statistics education community in developing assessment tools that measure various aspects of statistical proficiency. Currently, there are at least two large-scale programs aimed at designing instruments to assess statistics learning. One, based at Oklahoma University, is attempting to create a statistics concept inventory that will be similar to the force concept inventory used in physics education (Rhoads and Murphy, 2004). The other is the ARTIST project, based at the University of Minnesota, which maintains a database of assessment items designed to test statistical

literacy, thinking and reasoning (ARTIST). The work presented here might serve as a basis for modification of the work in progress. In general the process of creating and validating a framework is useful for providing improved measurement tools, and a common language for discussion and definition. It provides an empirical basis to findings and allows the linkage of research and assessment with curricular and policy decisions because it helps to understand what is being measured and what factors underlie the development of what is being measured (Murray, et al., 2005).

3.2 The Psychology Constructs: EU and NC

The quantitative analyses of the data collected in the Population and Exam Studies did not find many significant relationships between EU and NC and learning outcomes. Some relationships that might be investigated further using another sample are the relationship between EU and NC in college students, the relationship between EU, NC and GPA, the relationship between NC and retention in statistics and/or other university courses. As was discussed in the section above on limitations and delimitations, prior to any future the use of the EU and NC tasks with college students, the tasks should be validated for test-retest reliability with that population.

A possible study design to assess the retest-reliability of NC would be to test students in both calculus and statistics courses on NC at two times during the semester. The second testing should not be too close to a midterm or to the end of the semester so as not to bias the NC results. Students who feel overburdened with schoolwork may feel less inclined to expend cognitive energy so they may appear to have lower NC than they would at less stressful times in the semester. The NC data should be analyzed using a one-sample matched pairs t-test, and particular attention should be paid to the distribution

of the differences in scores. The data should also be analyzed for differences between calculus students and statistics students to assess whether the statistics students behave differently than the calculus students.

Testing the stability of EU categorization presents more of a challenge. The basis of this analysis might have one of two underlying assumption models. The first is that for short time periods, EU should be roughly stable, even in the population of undergraduates who are expected to be developing with regard to their epistemic outlook. The second possible assumption is that EU can develop rapidly in undergraduates so one would expect movement, mostly in the same direction, from many of the subjects. In order to assess relative stability of EU to decide which of the two models is more likely, a sample of undergraduates might be compared to a sample of adults, who are not likely to be developing with regard to epistemological understanding. If the proportion of undergraduates who switch categories is similar to the proportion of adults who switch categories, we might assume that EU is roughly stable over short time periods. We would then expect changes in both directions and could compare the data to that model. If it is found that the undergraduate sample is significantly more volatile that the adult sample, the undergraduate results would be compared to a model in which most subjects are expected to stay the same or move up one category and those in transition might be equally likely to test higher or lower on a given day.

3.3 Development of Statistical Proficiency

Two related delimitations of this study are, 1) the sample was a homogenous group of novices in statistics and 2) expert statisticians were not asked to react the interview and writing tasks. Therefore, this research does not provide results about the

development of statistical proficiency about hypothesis testing. A cross sectional study of students at many stages of statistical development using tasks similar to the interview and writing tasks developed for this work might provide a basis for a developmental model of understanding of hypothesis testing. In addition to the novice students who comprised the sample in this study, one could interview undergraduates who are not statistics majors, but who have taken more than one quantitative analysis course, undergraduate statistics majors, graduate students in statistics, practicing applied statisticians, and theoretical statisticians. The undergraduate and graduate statistics students could be sorted into categories based on the number of hours of statistics courses that had been completed. The results of such a study would provide information about both how proficiency develops and how expert proficiency in statistics differs from novice proficiency.

3.4 Psychology of Reasoning and the Statistics Classroom

Within the literature review of this work, I suggest several possible applications of the psychology on human reasoning to the statistics classroom, for example, belief bias, framing effects, and the effectiveness of frequency formats and distributional forms in problems involving conditional probability. Any of these topics could be the basis of an empirical study in statistics education. In this section I will discuss how the results of frequency formats and distributional forms studies might be used as a basis for study in a statistics class.

The original problem that provided a basis for findings about frequency formats and distributional forms in the psychology literature is the Medical Test Problem (Cosmides and Tooby, 1996), presented in Chapter 3 of this work. In my literature review, I cited the Testing for HIV problem (Moore, 2003), a problem from the statistics

textbook used at the University at which I conducted my study that is analogous to the Medical Test Problem. After the discussion of the findings in the psychology literature and based on those finding, I presented a version of the Testing for HIV problem that students should find easier to complete than the original version. To test whether the findings from psychology transfer directly to the domain of statistics learning, one could embed different versions of a problem, like the Testing for HIV problem, into a classroom assessment. In one possible design, each student would randomly receive one version of the problem and differences in performance between the groups could be assessed.

Manipulation of the question format is only one aspect of this literature that might be studied in the statistics classroom. Another finding from the psychology literature about these types of problems is that training subjects in the use of a representation, such as a grid or a Venn diagram, improves performance on conditional probability tasks (Cosmides and Tooby, 1996, Sedlmeier, 1999). Sedlmeier used as subjects in his study undergraduate students who had some background in statistics. He studied the growth in their performance in a laboratory setting. His work could be extended to students who are currently taking statistics by explicitly incorporating similar representations training for some students and then comparing the results of these students to those who did not receive training. As a matter of ethics, if representations training is found to have a significant positive effect on learning, the students who did not receive the training should have access to it at a later time.

In my literature review I claim that the findings on frequency format and distributional form, if they transfer to the classroom, have broader implications for

statistics learning. Some of the specific claims that I have made in the literature review are:

- When tests of proportions are introduced to students, the data should be presented initially in frequency format.
- There may be a benefit to presenting tests of proportions prior to tests of means.
- The distributional nature of sampling should be explicitly discussed when hypothesis testing is presented.

Each of these claims could serve as the basis for possible studies in statistics education. The suggestions for future research based on the psychological study of human reasoning are not limited to the particular domain of frequency format and distributional form. Studies analogous to those discussed in this section could be created to examine the application of other findings from psychology to the statistics classroom.

Appendix A – Epistemological Understanding Task

Directions: For each of the following, make one choice in part A. If you choose ii., in part A, make a choice in part B. Work quickly as your first thought is of interest.

Competing Statements: The number indicates the order in which the statements appeared on the instruments.

Statements about Aesthetic Judgments

1. Robin thinks the first piece of music they listen to is better.
   Chris thinks the second piece of music they listen to is better.

9. Robin thinks the first painting they look at is better.
   Chris thinks the second painting they look at is better.

11. Robin thinks the first book they read is better.
    Chris thinks the second book they read is better.

Statements about Value Judgments

2. Robin thinks lying is wrong.
   Chris thinks lying is permissible in certain situations.

5. Robin thinks the government should limit the number of children families are allowed to have to keep the population from getting too big.
   Chris thinks families should have as many children as they choose.

10. Robin thinks people should take responsibility for themselves.
    Chris thinks people should work together to take care of each other.

Statements about Truth in the Social World

3. Robin agrees with one book's explanation of how children learn language.
   Chris agrees with another book's explanation of how children learn language.

7. Robin has one view of why criminals keep going back to crime.
   Chris has a different view of why criminals keep going back to crime.

8. Robin thinks one book's explanation of why the Crimean wars began is right.
   Chris thinks another book's explanation of why the Crimean wars began is right.

Statements about Truth in the Physical World

4. Robin believes one mathematician's proof of the math formula is right.
   Chris believes another mathematician's proof of the math formula is right.

6. Robin agrees with one book's explanation of how the brain works.
   Chris agrees with another book's explanation of how the brain works.

12. Robin agrees with one book's explanation of what atoms are made up of.
    Chris agrees with another book's explanation of what atoms are made up of.

Response Choices:

A. Can only one of their views be right, or could both have some rightness? (CIRCLE ONE)

i. ONLY ONE RIGHT ii. BOTH COULD HAVE SOME RIGHTNESS

B. If both could be right, could one view be better or more right than the other? (CIRCLE ONE)

i. ONE COULD BE MORE RIGHT

ii. ONE COULD NOT BE MORE RIGHT THAN THE OTHER

Appendix B – Need for Cognition Task

Directions: For each of the statements, circle the response that best describes the strength of your agreement or disagreement with the statement. Work quickly as it is your first thought that is of interest.

Statements: Those with asterisks are reverse coded.

1. I would prefer complex to simple problems.
2. I like to have the responsibility of handling a situation that requires a lot of thinking.
3. *Thinking is not my idea of fun.
4. *I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
5. *I try to anticipate and avoid situations where there is likely chance I will have to think in depth about something,
6. I find satisfaction in deliberating hard for long hours.
7. *I only think as hard as I have to.
8. *I prefer to think about small, daily projects to long-term ones.
9. *I like tasks that require little thought once I've learned them.
10. The idea of relying on thought to make my way to the top appeals to me.
11. I really enjoy a task that involves coming up with new solutions to problems.
12. *Learning new ways to think doesn't excite me very much.
13. I prefer my life to be filled with puzzles that I must solve.
14. The notion of thinking abstractly is appealing to me.
15. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.
16. *I feel relief rather than satisfactions after completing a task that required a lot of mental effort.

17. *It's enough for me that something gets the job done; I don't care how or why it works.

18. I usually end up deliberating about issues even when they do not affect me personally.

Likert scale: The following choices appeared under each item, from left to right across the page.

Very Strongly Agree
Moderately Agree
Slightly Agree
Don't Know
Slightly Disagree
Moderately Disagree
Strongly Disagree
Very Strongly Disagree

Appendix C – Instrument Front Pages

1. PILOT STUDY

Survey Instrument IRB Protocol #2005-03-0094

1. Spring 05 M316 Unique Number _____

2. Gender: ____ M ____ F

3. Year in school: ___ Freshman
___ Sophomore
___ Junior
___ Senior

4. Cumulative UT GPA: _____ 5. Expected Grade in Statistics: _____

6. Last Math Course taken prior to M316:

_____ Semester and Year _____ Grade _____

9. Number of years of high school math _____

10. Last high school math class taken _____ Grade _____

11. My grades in mathematics tend to (mark an X is the space that best describes you experience)

_____ a. be too high compared to what I think I've learned.

_____ b. be too low compared to what I think I've learned.

_____ c. accurately reflect what I think I've learned.

12. If you took the SAT test, please record your scores:

SAT math _____

SAT verbal _____

13. One a scale of 1 to 10, where 10 is very sure and 1 means you have made a complete guess, how sure are you about the SAT scores you recorded? _____

14. If you did not take the SAT but know your score for another standardized test, (the ACT or TASP for example) record it here:

Test Name _____ Score _____

15. Is English your first language (circle one)? Yes No

16. What language to you speak at home? _____

2. POPULATION STUDY

IRB PROTOCOL # 2005-03-0094: There will be other opportunities to participate in this study. Subjects will be needed to retake this instrument in order to track changes in the qualities they measure over the course of the semester. Other subjects will be asked to answer statistics exercises and/or give interviews about their understandings of statistics. Some subjects will be offered payment or other compensation for their time.

If you would be interested in participating further in this study, please provide the following contact information. You will be contacted during the first week of December.

If you are not interested in participating further, do NOT fill out the contact information.

Yes, I am interested in participating further.

Name (printed): _____

Email address: _____

Phone: _____

Signature: _____

3. TARGETED STUDY

Survey Instrument IRB Protocol #2005-03-0094

1. Year in school: ___ Freshman
___ Sophomore
___ Junior
___ Senior

2. Cumulative UT GPA: _____

3. Grade in M316: _____

4. My grade in M316 was (mark an X in the space that best describes your experience)

_____ a. too high compared to what I think I've learned.

_____ b. too low compared to what I think I've learned.

_____ c. accurately reflect what I think I've learned.

Appendix D – Reading Comprehension Task: Pilot Study

Read the passage and then answer the questions that follow by marking an X in the space in front of the correct solution.

The determination of the sources of copper ore used in the manufacture of copper and bronze artifacts of Bronze Age civilizations would add greatly to our knowledge of cultural contacts and trade in that era. Researchers have analyzed artifacts and ores for their concentrations of elements, but for a variety of reasons, these studies have generally failed to provide evidence of the sources of copper used in all the objects. Elemental composition can vary within the same copper-ore lode, usually because of various admixtures of other elements, especially iron, lead, zinc, and arsenic. And high concentrations of cobalt or zinc, noticed in some artifacts, appear in a variety of copper-ore sources. Moreover, the processing of ores introduced poorly controlled changes in the concentrations of minor and trace elements in the resulting metal. Some elements evaporate during smelting and roasting; different temperatures and processes produce different degrees of loss. Finally, flux, which is sometimes added during smelting to remove waster material from the ore, could add quantities of elements to the final product.

An elemental property that is unchanged through these chemical processes is the isotopic composition of each metallic element in the ore. Isotopic composition, the percentages of the different isotopes of an element in a given sample of the element, is therefore particularly suitable as an indicator of the sources of the ore. Of course, for this purpose it is necessary to find an element whose isotopic composition is more or less constant throughout a given ore body, but varies from one copper ore body to another or, at least, from one geographic region to another.

The ideal choice, when isotopic composition is used to investigate the source of copper ore, would seem to be copper itself. It has been shown that small but measurable variations occur naturally in the isotopic composition of copper. However, the variations are large enough only in rare ores; between samples of the common ore minerals of copper, isotopic variations greater than the measurement error have not been found. An alternative choice is lead, which occurs in most copper and bronze artifacts of the Bronze Age in amounts consistent with the lead being derived from the copper ores and possibly from the fluxes. The isotopic composition of lead often varies from one source of common copper ore to another, with variations exceeding the measurement error; and preliminary studies indicate virtually uniform isotopic composition of the lead from a single copper-ore source. While some of the lead found in an artifact may have been introduced from flux or when other metals were added to the copper ore, lead so added in Bronze Age processing would usually have the same isotopic composition as the lead in

the copper ore. Lead isotope studies may this prove useful for interpreting the archaeological record of the Bronze Age.

1. The primary purpose of this passage is to

_____ a. discuss the techniques of analyzing lead isotope composition.
_____ b. propose a way to determine the origin of the copper in certain artifacts.
_____ c. resolve a dispute concerning the analysis of copper ore.
_____ d. describe the deficiencies of a currently used method of chemical analysis of certain metals.
_____ e. offer an interpretation of the archaeological record of the Bronze Age.

2. The author first mentions the additions of flux during smelting in order to

_____ a. give a reason for the failure of elemental composition studies to determine ore sources.
_____ b. illustrate differences between various Bronze Age civilizations.
_____ c. show the need for using high smelting temperatures.
_____ d. illustrate the uniformity of lead isotope composition.
_____ e. explain the success of copper isotope composition analysis.

3. The author suggests which of the following about a Bronze Age artifact containing high concentrations of cobalt or zinc?

_____ a. It could not be reliably tested for its elemental composition.
_____ b. It could not be reliably tested for its copper isotope composition.
_____ c. It could not be reliably tested for its lead isotope composition.
_____ d. It could have been manufactured from ore from any one of a variety of sources.
_____ e. It could have been produced by the addition of other metals during the processing of the copper ore.

4. According to the passage, possible sources of the lead found in a copper or bronze artifact include which of the following?

I. The copper ore used to manufacture the artifact
II. Flux added during the processing of the copper ore
III. Other metal added during processing of the copper ore

_____ a. I only
_____ b. II only
_____ c. III only
_____ d. II and III only
_____ e. I, II, and III

5. The author rejects copper as the "ideal choice" mentioned in the first sentence of the third paragraph because

_____ a. the concentration of copper in Bronze Age artifacts varies.
_____ b. elements other than copper may be introduced during smelting.
_____ c. the isotopic composition of copper changes during smelting.
_____ d. among common copper ores, differences in copper isotope composition are too small.
_____ e. within a single source of copper ore, copper isotope composition can very substantially.

6. The author makes which of the following statements about lead isotope composition?

_____ a. It often varies from one copper-ore source to another.
_____ b. It sometimes varies over short distances in a single copper-ore sources.
_____ c. It can vary during the testing of artifacts, producing measurement error.
_____ d. It frequently changes during smelting and roasting.
_____ e. It may change when artifacts are buried for thousands of years.

7. It can be inferred from the passage that the use of flux in processing copper ore can alter the lead isotope composition of the resulting metal EXCEPT when

_____ a. there is a smaller concentration of lead in the flux than in the copper ore.
_____ b. the concentration of lead in the flux is equivalent to that of the lead in the ore.
_____ c. some of the lead in the flux is equivalent to that of the lead in the ore.
_____ d. any lead in the flux has the same isotopic composition as the lead in the ore.
_____ e. other metals are added during processing.

Appendix E – Statistics Tasks

1. PILOT STUDY

1.1 Long Answer Assessment of Procedural Fluency

A recent CNN/Gallup poll asked a random sample of 1127 registered voters if they approved of the job that George W. Bush is doing as president. 527 of those polled responded that they approved of the job that George W. Bush is doing as president. Do the results of this survey provide evidence that George W. Bush's true approval rating is less than 50%? Justify your answer with a complete hypothesis test.

1.2 Multiple-Choice Items Assessing Conceptual Understanding and Statistical Reasoning

1.2.1 Multiple Choice Items that Assess Conceptual Understanding

2. A researcher tested the null hypothesis that two population means are equal (H_o: $\mu_1 = \mu_2$) against a two-sided alternative. A t-test produced $p = 0.01$. You may assume that all assumptions of the test have been satisfied. Mark each statement about the p-value as true or false by circling the appropriate response.

   A. There is a 1% chance of getting a result this extreme or even more extreme than the observed one when H_o is true.
   True False

   B. There is a 1% likelihood that the result happened by chance.
   True False

   C. There is a 1% chance that the null hypothesis is true.

   True False

   D. There is a 1% chance that if the research redid his study he would get the same results.
   True False

   E. There is a 1% chance that the decision to reject H_o is wrong.
   True False

12. It is reported that scores on a particular test of historical trivia given to high school students are approximately normally distributed with a mean of 85. Mrs. Rose believes that her 5 classes of high school seniors will score significantly better than the national average on this test. At the end of the semester, Mrs. Rose administers the historical trivia test to her students. The students score an average of 89 on this test. After conducting the appropriate statistical test, Mrs. Rose finds that the *p*-value is .025. Mark each statement as true or false.

A. There is a 2.5% chance that if Mrs. Rose gave another version of the test to her students she would get the same results.
True False

B. There is a 2.5% likelihood that Mrs. Rose's students good performance was due to chance.

True False

C. There is a 2.5% chance that Mrs. Rose's students have the same capabilities as all students across the nation.
True False

D. There is a 2.5% chance that Mrs. Rose's students would score this high or higher if they actually had the same capabilities of all students across the nation.
True False

4. Your friend buys a carton (10 packages) of cigarettes and repeats the study you did in question 3. She receives the same sample mean that you did in your study with only 1 package of cigarettes. Which of the following statements is true? (Mark one choice from each list in the space BEFORE the choice.)

List 1	List 2
_____ Her p-value will be larger.	_____ She will find the same level of evidence against H_o.
_____ Her p-value will be smaller.	_____ She will find more evidence against H_0.
_____ Her p-value will be the same.	_____ She will find less evidence against H_0.

6. Imagine you have a huge jar of marbles with many different colors. We know that 35% of the marbles are orange. Five students each take a random sample of 20 marbles, one at a time. Which sequence below seems most plausible for the percent of orange marbles obtained in these five samples?

_____ a. 30%, 35%, 15%, 40%, 50%.

_____ b. 35%, 35%, 35%, 35%, 35%.

_____ c. 5%, 60%, 10%, 50%, 95%.

9. A study was planned to examine the length of a certain species of fish from one lake. The plan was to take a random sample of 100 fish and examine the results. Numerical summaries on lengths of the fish measured in this study are given.

Mean	26.8
Median	29.4
Standard Deviation	5.0
Min	12.7
Max	33.40

Another researcher took a different random sample of 100 fish from the same lake and found the mean of this sample to be 25.3 inches. Should the researcher be surprised by this result? (select the best answer)

_____ a. No, because the 25.3 is less than one standard deviation from the mean of the first researcher's sample (26.8).

_____ b. Yes, because with a large sample of 100 fish the two sample means should be almost identical.

_____ c. Yes, because the difference between 25.3 and 26.8 is much larger than the expected sampling error.

10. A certain manufacturer claims that they produce 50% brown candies. Sam plans to buy a large family size bag of these candies and Kerry plans to buy a small fun size bag. Which bag is more likely to have more than 70% brown candies?

_____ a. Sam, because there are more candies, so his bag can have more brown candies.

_____ b. Sam, because there is more variability in the proportion of browns among larger samples.

_____ c. Kerry, because there is more variability in the proportion of browns among smaller samples.

_____ d. Kerry, because most small bags will have more than 50% brown candies.

_____ e. Both have the same chance because they are both random samples.

1.2.2 Multiple Choice Items that Assess Statistical Reasoning

1. A computer technician uses an instrument to test whether or not motherboard is defective. The instrument sometimes fails to detect a defective motherboard. The technician will try to replace a motherboard only if the testing instrument indicates that it is defective. The hypothesis he is testing is that the motherboard is not defective. An alternative hypothesis is that the motherboard is defective.

   A. If the technician rejects the null hypothesis, which of the following statements is true?

      _____ a. The motherboard is definitely defective and needs to be repaired.

      _____ b. The technician decides that the motherboard is defective, but it could be good.

      _____ c. None of the above.

   B. If the technician does not reject the null hypothesis, which of the following statements is true?

      _____ a. The motherboard is definitely good and does not need to be repaired.

      _____ b. The motherboard is most likely good, but it could be defective.

      _____ c. None of the above.

3. The nicotine content in cigarettes of a certain brand is normally distributed with mean (in milligrams) μ and standard deviation $\sigma = 0.1$. The brand advertises that the mean nicotine content of their cigarettes is 1.5 mg, but you believe that the mean nicotine content is actually higher than advertised. To explore this, you test the hypotheses:

   H_0: $\mu = 1.5$
   H_a: $\mu > 1.5$

   using one package of cigarettes and you obtain a *P*-value of 0.072. Which of the following is true?

   _____ a. At the $\alpha = 0.05$ significance level, you have proven that H_0 is true.

   _____ b. There is some evidence against H_0, and a study using a larger sample size may be worthwhile.

   _____ c. This should be viewed as a pilot study and the data suggests that further investigation of the hypotheses will not be fruitful at the $\alpha = 0.05$ significance level.

5. Researchers surveyed 1,000 randomly selected adults in the U.S. A statistically significant, strong positive correlation was found between income level and the number of containers of recycling they typically collect in a week. Based on this information alone, select the best interpretation of the data.

_____ a. Earning more money allows people to recycle more than people who earn less money.

_____ b. This sample is too small to draw any conclusions about the relationship between income level and amount of recycling for adults in the U.S.

_____ c. A conclusion that earning more money causes more recycling among U.S. adults is not well supported.

7. A researcher in environmental science is conducting a study on armadillos. His goal is to compare armadillos that have been exposed to a particular pesticide to those that have not. He has 4 armadillos in each treatment group. The exposed armadillos showed higher levels of the indicator enzyme. However, a test of significance was correctly conducted and showed no significant difference in average enzyme level between the armadillos that were exposed to the pesticide and those that were not. What conclusion can the graduate student make from these results?

_____ a. The researcher must have made a mistake in his calculations; there should be a significant difference.

_____ b. The sample size may be too small to detect a statistically significant difference.

_____ c. It must be true that the pesticide does not cause higher levels of the enzyme.

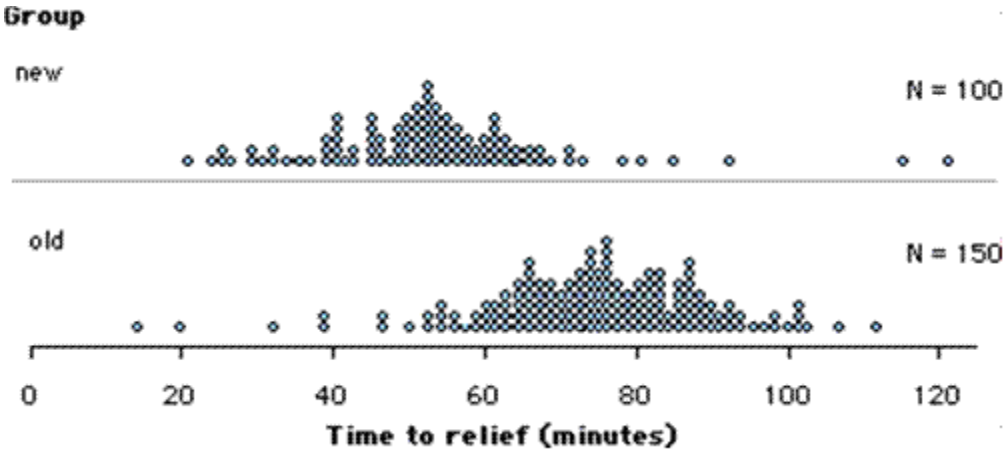
8. A newspaper article stated that the Austin City Council received 812 letters from around Travis County about the proposed Austin area toll roads. Of these 812 letters, 800 expressed the opinion that the new roads should not be toll roads. A statistics student was going to use this sample information to conduct a test of significance of whether more than 95% of all Travis County residents feel that the new roads should not be toll roads. What would you tell this student?

_____ a. This is a large enough sample to provide an accurate estimate of the opinion of Travis County residents on the issue.

_____ b. The necessary conditions for a test of significance are not satisfied, so no statistical test is appropriate.

_____ c. With such a large number of people opposing the toll roads, there is no need for a statistical test.

11. A drug company developed a new formula for their headache medication. To test the effectiveness of this new formula, they gave it to 100 people with headaches and timed how many minutes it took for the patient to no longer have a headache. They compared the result from this test to previous results from 150 patients using the old formula under the exact same conditions. The results from both of these clinical trials are shown below. Below are statements made by three statistics students. For each statement, indicate whether you think the student's conclusion is valid.



A. The old formula works better. Two people who took the old formula felt relief in less than 20 minutes, compared to none who took the new formula. Also, the worst result - near 120 minutes - was with the new formula.

Valid. Not valid.

B. The average time for the new formula is lower than the average time for the old formula. I'd conclude that people taking the new formula will tend to feel relief about 20 minutes sooner than those taking the old formula.

Valid. Not valid.

C. I would not conclude anything from these data. The number of patients in the two groups is not the same so there is no fair way to compare the two formulas.

Valid. Not valid.

13. Your statistics instructor bought a large bag of peanut butter M&M's (her favorite) and counted 11 green M&M's out of 390 in the bag. The M&M/Mars website (www.mms.com) reports that 20% of peanut butter M&Ms are green. In a hypothesis this test of the null hypothesis, 20% of peanut butter M&M's are green, against a two-sided alternative, the test statistics was z = -8.46 and the p-value nearly zero.

    Below is a list of possible reasons for this outcome. Rank the most plausible as 1 and the second most plausible as 2.

    _____ a. The M&M/Mars website is probably correct; your instructor chose an unlucky sample.
    _____ b. M&M/Mars has an error on their website; the proportion of M&M's that are green is not 20%.
    _____ c. The M&M/Mars website is probably correct; the machine that mixes the different M&M's colors together does not do a very good job so one bag is not a random sample of M&M's.
    _____ d. It is impossible to tell whether the M&M/Mars website is correct from this experiment; one bag of M&M's is too small a sample to produce reliable results so your instructor should redo the experiment using more bags of M&M's.
    _____ e. The M&M/Mars website is probably correct in general, but on that day the factory might have been out of green dye.

1.3 Open-Ended Assessment of Strategic Competence and Statistical Reasoning

Imagine you receive the following email from your dad:

Hey Kiddo,

I'm worried about Grandma. Remember that she was diagnosed with high blood pressure? Well, she's currently taking the medication Makemewell to lower her blood pressure. At the time of Grandma's diagnosis, her doctor said that a randomized, double-blind experiment had been conducted and that Makemewell was shown more effective in lowering blood pressure than a placebo. He said that the 40 high blood pressure patients who took Makemewell had an average drop in systolic blood pressure of 19.5 millimeters of mercury (mm Hg). The 40 high blood pressure patients who took the placebo had an average drop of 9.1 mm Hg. The doctor said that these results were significant with $p < .001$.

To tell you the honest truth, I don't understand much of what the doctor said, but I believed and trusted the doctor so Grandma started on the medicine. Now I've heard two stories that make me think differently. Larry, our next-door neighbor, was taking Makemewell and he got a terrible fever that put him in the hospital. Also, my co-

worker, Sally, actually had her blood pressure go up while she was taking Makemewell. I am now very suspicious of this medication and I am thinking of taking your grandmother off of it.

I know that you are taking a statistics course. Based on what I've written, could you tell me whether you think grandma should stay on this medication? Please explain fully how you came to this conclusion and don't use any technical language. Also, would you be sure to explain what you think the doctor meant by his description of the Makemewell study? I would like to use what you write as the basis for a follow-up conversation with the doctor.

Thanks and I've enclosed a check to get you through the rest of the semester.

Dad.

Use this page to write a response to your "father's" email.

Choose the statement that best describes the email you wrote in the previous page.

_____ a. It is the response I thought would earn me high marks from my statistics instructor.

_____ b. It is the response I would have really written if I had really gotten that email from my dad about my grandmother.

_____ c. There would have been no difference between a response to my dad and a response for a statistics class.

Other: (Please explain)

If you chose option A above, please us the space below to explain how your answer would have differed if you were writing for option B.

If you chose option B above, please us the space below to explain how your answer would have differed if you were writing for option A.

2. EXAM STUDY

2.1 Multiple Choice Questions

1. The nicotine content in cigarettes of a certain brand is normally distributed with mean (in milligrams) μ and standard deviation $\sigma = 0.1$. The brand advertises that the mean nicotine content of their cigarettes is 1.5 mg, but you believe that the mean nicotine content is actually higher than advertised. To explore this, you test the hypotheses:

   H_0: $\mu = 1.5$
   H_a: $\mu > 1.5$

   using one package of cigarettes and you obtain a *P*-value of 0.072. Consider the following statements:

   I. There is a 7.2% chance that if you tested another package of cigarettes she would get the same results.
   II. There is a 7.2% chance that the mean nicotine content for this brand of cigarettes is 1.5 mg.
   III. There is a 7.2% chance that you would have obtained the mean that you did or one even larger from your sample if mean nicotine content for this cigarette brand really was 1.5 mg.

   a. All of the statements are false.
   b. Only statement I is true.
   c. Only statement II is true.
   d. Only statement III is true.
   e. Statements II and III are both true.

2. Your friend buys a carton (10 packages) of cigarettes and repeats the study you did in question 1. She receives the same sample mean that you did in your study. Which of the following statements is true?

   a. Her p-value will be larger and she will find more evidence against H_0.
   b. Her p-value will be smaller and she will find more evidence against H_0.
   c. Her p-value will be smaller and she will find less evidence against H_0.
   d. Her p-value will be the same and she will find the same level of evidence against H_o.
   e. Her p-value will be the same and she will find more evidence against H_0.

3. A certain manufacturer claims that they produce 50% brown candies. Sam plans to buy a large family size bag of these candies and Kerry plans to buy a small fun size bag. Which bag is more likely to have more than 70% brown candies?

    a. Sam, because there are more total candies in his bag, so his bag can have more brown candies.
    b. Sam, because there is more variability in the proportion of browns among larger samples.
    c. Kerry, because there is more variability in the proportion of browns among smaller samples.
    d. Kerry, because most small bags will have more than 50% brown candies.
    e. Both have the same chance because they are both random samples.

2.2 Long Answer Hypothesis Test Questions

One Sample Tests of Proportion:

A recent CNN/Gallup poll asked a random sample of 1127 registered voters if they approved of the job that George W. Bush is doing as president. 527 of those polled responded that they approved of the job that George W. Bush is doing as president. Do the results of this survey provide evidence that George W. Bush's true approval rating is less than 50%? Justify your answer with a complete hypothesis test.

The company Mason-Dixon Polling and Research, Inc. conducted an opinion poll for the St. Paul Pioneer Press and Minnesota Public Radio from Oct. 30 through Nov. 1, 2002, just prior to the election on Nov. 5, 2002. The poll surveyed 625 potential voters. One of the questions asked if the person thought Walter Mondale was the best choice to replace Paul Wellstone as the Democratic candidate for Minnesota senator. Of the 625 people, 344 answered YES to this question. Does the evidence from this sample support the hypothesis that more than half of all potential voters thought Mondale was the best choice? Justify your answer with a complete hypothesis test.

The experience of unrequited love is virtually universal at some point in life. It is believed that only 5% of adults have never experienced it. Some social psychologists think the percentage is even lower. Dr. Baumeister and Sara Wotman, a graduate student, found in a study of 155 men and women that only about 2 percent had never loved someone who rejected them, or found themselves the object of romantic passion that they did not reciprocate. Is Dr. Baumeister able to refute the claim that 5% of adults have never experienced unrequited love? Justify your response with a complete hypothesis test.

One Sample Test of Means:

The American College Testing Corporation has a test called the ACT that is used for admissions to college. The American College Testing Corporation reports that students who attend 2-year colleges in the Midwest United States have an average ACT score of 21.5. A random sample of 51 students who started at a Minneapolis Community College in fall 2002 had an average ACT score of 20.4 with a sample standard deviation of 3.17. Use a hypothesis test to determine if the average ACT score for all students who attend this Minneapolis Community College is significantly different from the ACT scores of students who attend two-year colleges.

A newspaper article claims that the average age for people on food stamps is 40 years. You believe that the average age is less than that. You take a random sample of 40 people who receive food stamps, and find their average age to be 39.2 years, with a standard deviation of 5.2 years. Do you have enough evidence to say the newspaper's claim is incorrect? Justify your answer with a complete hypothesis test. You may assume that the technical conditions are met.

One Sample Matched Pairs Design:

Dr. Hanson's mother Sally and her sister Pat have a rivalry about who makes the better apple pie. Sally and Pat make pies and select 16 people at random from their church eat a piece of pie and rate them on a scale of 1-10. Subjects tasted the pies in random order. The ratings are given in the following table. Each column represents the rating by one taster.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Sally	6	8	3	9	9	9	8	8	6	7	8	7	6	7	6	8
Pat	7	7	3	9	9	8	6	6	6	7	4	8	6	10	2	7

a. Perform an appropriate hypothesis test to see if one sister's pies are significantly better than the others.
b. Sketch the data and a comment on the validity of the conclusions.

A sports writer wished to see if a football filled with helium travels farther when kicked, on average, than a football filled with air. To test this, the writer used 12 adult male volunteers. Each subject kicked each of the two footballs, one filled with helium and one filled with air, once. The order of the balls was randomized. The table below gives the lengths of the kicks in yards.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Air	25	26	33	22	22	27	28	34	26	15	32	25
Helium	25	25	35	30	39	28	14	22	32	29	29	29

a. Perform an appropriate hypothesis test to test the sportswriter's claim.
b. Sketch the data and a comment on the validity of the conclusions.

Two Sample Test for Means:

Texans claim that everything is bigger in Texas. In order to test that notion, a researcher in Oklahoma selected a random sample of baseball players from the Texas and Oklahoma baseball teams. The weights, in pounds, of 8 Oklahoma players and 9 Texas players are given in the table below.

Oklahoma	Texas
180	193
185	210
178	199
195	190
200	190
180	180
170	185
160	167
	205

a. Use the data to test the hypothesis that Texas athletes are heavier than Oklahoma athletes.
b. Use the stem below to create a back-to-back stem plot of your data and assess the validity of the conclusions of your hypothesis test in part a.

| 16 |
| 17 |
| 18 |
| 19 |
| 20 |
| 21 |

A researcher wanted to test whether people who exercise regularly tend to have lower resting heart rates than people who do not exercise regularly. She tested the resting heart rates, in beats per minute, of 18 people, 10 who exercised regularly and 8 who did not. The data is in the table.

Exercise	No Exercise
62	72
72	84
59	66
63	72
75	62
60	84
52	76
80	60
68	
64	

a. Use the data to test the hypothesis that people who exercise have lower resting heart rates than people who do not exercise regularly.

b. Use the stem below to create a back-to-back stem plot of your data and assess the validity of the conclusions of your hypothesis test in part a.

| 5 |
| 6 |
| 7 |
| 8 |

Appendix F – Interview Protocol: Targeted Study

1. SCENARIO A: EMAIL FROM DAD

1.1 Written Task

Imagine you receive the following email from your dad:

> Hey Kiddo,
>
> I'm worried about Grandma. Remember that she was diagnosed with high blood pressure? Well, she's currently taking an experimental medication called Makemewell to lower her blood pressure. At the time of Grandma's diagnosis, her doctor said that a randomized, double blind experiment had been conducted and that Makemewell was shown more effective in lowering blood pressure than a placebo. He said that the 40 high blood pressure patients who took Makemewell had an average drop in systolic blood pressure of 19.5 millimeters of mercury (mm Hg). The 40 high blood pressure patients who took the placebo had an average drop of 9.1 mm Hg. The doctor said that these results were significant with $p < .001$.
>
> To tell you the honest truth, I don't understand much of what the doctor said, but I believed and trusted the doctor so Grandma started on the medicine. Now I've heard two stories that make me think differently. Larry, our next-door neighbor, was taking Makemewell and he got a terrible fever that put him in the hospital. Also, my co-worker, Sally, actually had her blood pressure go up while she was taking Makemewell. I am now very suspicious of this medication and I am thinking of taking your grandmother off of it.
>
> I know that you are taking a statistics course. Based on what I've written, could you tell me whether you think grandma should stay on this medication? Please explain fully how you came to this conclusion and don't use any technical language. Also, would you be sure to explain what you think the doctor meant by his description of the Makemewell study? I would like to use what you write as the basis for a follow-up conversation with the doctor.
>
> Thanks and I've enclosed a check to get you through the rest of the semester.
>
> Dad.

Use the paper provided to draft a response to this email.

1.2 Interview Protocol

Before discussing the email:

1. Would you ever receive this type of email?
2. Are any of your grandparents alive?
3. Are you close to them?
4. Do they have medical conditions?
5. What do you think of doctors in general?
6. Have you ever read about pharmaceutical trials? What do you think about the drug companies/drug testing?

Discussion of email:

1. Let's role play. Pretend that I am your Dad (or Mom) and I've called you to ask you about your response to my email.

   [This set of questions is designed to probe the written responses. Any of the questions can be prefaced with, "I think you said this in the email, but I'm not sure I understood correctly" or "You didn't mention this, but I'm curious…"]

   a. What do you think I should do about keeping Grandma on the medication?
   b. Why do you think I should take Grandma off (leave her on) Makemewell?
   c. What did the doctor mean by "randomized, double blind"?
   d. Why does randomized or double blind matter?
   e. What did the doctor mean by $p < .001$?

2. If subject mentions possible side effects, say, "Oh, yea, the doctor said something about that. I wrote it down but I didn't realize it was important. He said that the incidence of side effects in the two groups was not significant, $p = .13$. Can you tell me what that means?"
3. Probe for incorporation of anecdotal evidence. If subject brings it up, probe subject's thinking about the anecdotal evidence. If subject does not bring it up, ask "Well, what about the stories from Larry and Sally? Should we be worried about them?"
4. Finish the interview by asking "So, let me make sure I understand you. You think Grandma should stay on (go off) the medication?" Can you tell me your reasons one more time? I want to have them written down for when I talk to your uncle (the doctor) again."

2. SCENARIO B: PARANORMAL PSYCHOLOGY

2.1 Interview Protocol

Before giving the article:

1. Do you believe in ESP – extra sensory perception?
2. Have you ever met a psychic?
3. Have you had your fortune told or your cards read?
4. Do you believe that some people can see the future?

After reading the article:

1. What do you think of the article?
2. Does reading the article change your beliefs about ESP? How? Why? Why not?
3. Can you explain to me the results the article talks about? The z and the p-value?
4. What do they mean when they say the results were statistically significant?
5. Do you have any comments about the way they are testing whether people have ESP?
6. What did you think about the way the study was done? Can you think of any improvements? Follow up questions or studies?

For people who think the article is not convincing:

7. a. If the article doesn't convince you about the existence of ESP, to what do you attribute the "significant" results?
   b. What if I told you that the results of this study have been replicated by two other sets of researchers? In one of the replications, there were 120 trials with dynamic targets and the p-value was .02 and in the other there were 185 trials and the p-value was .03. Does that convince you of the existence of ESP?
   b. Is there any kind of study or result that would convince you that there are people who do have ESP?

For subjects who are convinced by the article:

8. What if I told you that after this article was published other research groups tried to replicate the results? In fact, ten sets of researchers have done the same experiment and none have achieved statistical significance. They all had over 100 trials and none had a p-value lower than .14. Two had p-values over .5.
   a. Does that change your perception of this study? ESP?
   b. How (or Why not?)
9. So, tell me again after reading the article, what is your opinion about ESP? And why do you believe that?

2.2 Reading Task

Does ESP exist?

Extrasensory perception (ESP) is the apparent ability to obtain information in ways that exclude ordinary sensory channels. Early ESP research focused on having people try to guess what symbol was on the other side of a faced-down card, to see if they could guess at a better rate than would be expected by chance. In recent years, experimenters have used more interesting targets, like photographs, outdoor scenes, or short movie segments.

In 1994, two psychology researchers, Bem and Honorton, published an account of their 6 and a half year study of ESP in the journal, *Psychological Bulletin.*

In all of the Bem and Honorton studies, a receiver and sender are isolated in separate chambers that are insulated for sound. The receiver begins the session with a 14-min period of progressive relaxation. After that the receiver goes into a sensory deprivation state that is called "ganzfeld stimulation." During the "ganzfeld," the receiver has ping-pong ball halves taped over the eyes and headphones over the ears; a red floodlight directed toward the eyes produces a constant and unchanging visual field, and white noise is played through the headphones. During the 30-minute ganzfeld process, the receiver describes his or her thoughts and images aloud. While the receiver experiences the ganzfeld process, the sender concentrates on a randomly selected target. At the end of the ganzfeld period, the receiver is shown four stimuli. Without knowing which of the four had been the target, the receiver rates each stimulus for its similarity to his or her thoughts during the ganzfeld period. A trial is labeled "successful" if the target stimulus was stimulus that was rated the highest by the receiver.

A computer chose the all of the stimuli at random, both the target stimulus and the set of stimuli to be rated. The experimenters did not know the identity of the target stimulus until after the experiment was concluded. There were 80 still pictures (static targets) and 80 short video segments (dynamic targets) from which the computer would choose. The static targets included art prints, photographs, and magazine advertisements; the dynamic targets included excerpts of approximately 1-min duration from motion pictures, TV shows, and cartoons.

When the sender was looking at a "static" image, he saw the same image for the entire time on the television screen. If the sender was viewing a "dynamic" image, the short video clip played repeatedly during the experiment. When the receiver was rating the four images by how closely they resembled the thoughts that occurred during ganzfeld, all four images were of the same type (static or dynamic) to eliminate biases due to preference for one type of picture.

Altogether, 100 men and 140 women participated as receivers in 354 sessions during the research program. The experimental program included eleven studies. Eight separate experimenters conducted the studies. Some of the studies used receivers who were novice (first time) participants. Other studies used experienced participants.

One of the findings reported in the Bem and Honorton paper was that subjects are more likely to
exhibit ESP when the target is dynamic rather than static. Results from the experiment are shown in the table below. Because the right answer is one of four possibilities, receivers who are

Results of the Bem and Honorton ESP Study

	Successful Guess?			
	Yes	No	Total	Percent Success
Static Picture	45	119	164	27.4%
Dynamic Picture	61	104	165	37.0%

guessing randomly will be correct about 25% of the time. For the static pictures, the 27% rate is about what is expected from random guessing. The results of a one sided, one-sample test for proportions are not statistically significant ($z = .72$ and p-value = .235) for this case. The results for the dynamic pictures are statistically significant ($z = 3.55$, p-value = .0002). A two-sided, two-sample test for proportions indicates that there is some evidence against the hypothesis that the two types of pictures, static and dynamic, produce the same ESP results ($z = -1.85$, p-value = .0644).

3. SCENARIO C: LYON DIET STUDY

3.1 Interview Protocol

Before giving the article:

1. Can you tell me a little about your diet?
   a. Do you cook or eat out?
   b. Where do you usually eat?
   c. What do you typically eat for breakfast? Lunch? Dinner?
   d. Are you careful about what you eat? Why (or why not?)
   e. Do you think a lot about your diet?
2. Do you worry about your health at all?
3. Do you think about or plan for the future?

After reading the article:

1. What do you think of the article?
2. Based on the article, would you consider changing your diet? Why? Why Not?
3. Would you consider talking to your parents about changing their diet? Why? Why not?
4. If no: What kind of evidence would make you think about changing your diet?
5. If yes: How would you describe the study and results to someone if you wanted to convince them to change their diet?
6. What did you think about the way the study was done? Can you think of any improvements? Follow up questions or studies?
7. Have you read other articles like this one that link diet or exercise to reducing cancer risk? How did this study compare to others you've read?

3.2 Reading Task

Possible Health Benefits of a Mediterranean Style Diet

What is the Lyon Diet Heart Study?

This was a randomized, controlled trial with human subjects. Its goal was to test the effectiveness of a Mediterranean-type diet on the rate of coronary events in people who have had a first heart attack. The results suggest that a Mediterranean-style diet may help reduce recurrent events in patients with heart disease.

What were the methods used?

A total of 302 experimental- and 303 control-group subjects were randomized into the study. All were patients who had survived a first heart attack. The two groups had a similar coronary risk factor profile (blood lipids and lipoproteins, systolic and diastolic blood pressure, body mass index and smoking status). Patients in the experimental group were asked to comply with a specific Mediterranean-type diet. Patients in the control group received no dietary advice from the researchers but were asked by their physicians to follow a prudent diet.

An intermediate analysis was performed after a minimum follow-up of one year for each patient. The study was stopped at that point because of significant beneficial effects noted in the original group.

What is a Mediterranean-style diet?

While there is no one, typical, "Mediterranean" diet, the common dietary pattern of the 16 countries that border the Mediterranean has these characteristics:

* high in fruits, vegetables, bread and other cereals, potatoes, beans, nuts and seeds
* includes olive oil as an important source of monounsaturated fat
* dairy products, fish and poultry consumed in low to moderate amounts, little red meat
* eggs consumed zero to four times weekly
* wine consumed in low to moderate amounts

The Mediterranean-style diet used in the Lyon Diet Heart Study was quite comparable to the common pattern but it specifically required participants to eat foods high in alpha-linolenic (lin"o-LEN'ik) acid (a type of polyunsaturated omega-3 fatty acid). It included

* more bread, more root vegetables and green vegetables, more fish.
* less beef, lamb and pork (replaced with poultry).

* no day without fruit.
* butter and cream were replaced with margarine high in alpha-linolenic acid.

The diet averaged 30 percent of calories from fat, 8 percent from saturated fat, 13 percent from monounsaturated fat, 5 percent from polyunsaturated fat and 203 mg/day of cholesterol.

What was the diet in the control group?

People in the control group consumed a diet with about 34 percent of calories from fat, 12 percent from saturated fat, 11 percent from monounsaturated fat, 6 percent polyunsaturated fat and 312 mg/day of cholesterol. This diet is comparable to what is typically consumed in the United States.

What were the results of the study?

There were a total of 38 deaths (24 in controls vs. 14 in the experimental group), including 25 cardiac deaths (19 vs. 6) and 7 cancer deaths (4 vs. 3), and 24 cancers (17 vs. 7). Exclusion of early cancer diagnoses (within the first 24 months after entry into the trial) left a total of 14 cancers (12 vs. 2).

Statistical results were computed after adjustment for age, sex, smoking, leukocyte count, cholesterol level, and aspirin use.

The experimental subjects compared with control subjects were less likely to have died during the study ($p = .03$) and were less likely to have developed cancer ($p = .05$)

This randomized trial suggests that cardiac patients following a Mediterranean diet have a prolonged survival and may also be protected against cancer.

Appendix F – Summaries of the Interview Transcripts

Each of the following interview summaries has the same structure. The first paragraph of each summary describes the educational background of the subject as well as the EU and NC categorization and any situational features that may have interacted with either the interview, learning of statistics, or EU and NC categorization. The second paragraph describes the pre-existing beliefs, knowledge and experiences that may affect how the subject will interact with the interview tasks. These are provided as context for the reader. The penultimate paragraph describes the subject's overall treatment of the subject of the p-value. The final paragraph describes the subject's overall treatment of experimental design issues. These paragraphs provide the basis of two of the discussions presented in the next chapter.

The middle paragraphs of each summary describe the subject's reactions to the individual tasks. In the summary of the Email from Dad task, all elements that were assessed in the analysis of the writing samples are addressed. As the summaries were constructed, themes specific to each of the additional tasks emerged. In the discussions of the Lyon Diet Study task subjects tend to compare of the task article to other diet articles they have read and to give opinions about the reaction of the general public to data presented statistically versus anecdotal evidence. The discussions of the ESP Study tend to include a reaction to evidence that supports a conclusion in which the subjects do not believe. These reactions include meta-cognitive statements made by the subjects about the contradictory nature of their reactions to the studies and are included in the summaries. The common themes that result from these discussions are also discussed in the next chapter.

The names of the subjects have been changed. Quotes are used in the summaries to indicate that the subject used the exact word or phrase. The order of the ESP Study and Lyon Diet Study tasks was randomly chosen for each subject. The order of presentation in the summary is the same as the order in which the subjects completed the tasks. It should be noted that the EU-NC categories listed below were created using each subject's OEU classification. One of the variables under consideration, however, is PEU classification. All of the HighNC subjects in the Targeted Study were Evaluativists when considering judgments about the physical world so they would reside in the same NC-PEU categories. Each of the other five subjects represents a single NC by PEU category.

1. THE HIGHNC-EVALUATIVISTS

1.1 Bradford

Bradford is a 19-year-old, White sophomore pre-med student, intending to complete a B.S. in psychology with biology minor. He has a 4.0 G.PA. and reports that the A he earned in statistics accurately reflects his learning. Bradford tested as a HighNC-Evaluativist in both September and February. Bradford's opinion of doctors is positive, especially given that he would like to be one someday. Bradford's father, who is an engineer, is unlikely to ask Bradford for help with statistics. Bradford is close to his remaining grandparents, one of who has had a heart transplant. He tries to maintain a good diet because he is currently healthy and would like to stay that way. He knows that eating well is good for his health. He finds eating healthy difficult given the limited options in the cafeteria. Bradford's planning for the future seems limited to assessing how his C.V. will look to medical schools when he applies for admission in three years. He does not believe in ESP.

Bradford suggests that Grandma stay on the medication. He says that Sally represents variation; no drug is a miracle drug that works for everyone. About Larry, he tells his father that there was nothing in his email that indicated that the fever was caused by the medication and, if there were serious side effects, the drug would not have been approved by the F.D.A. When told that the results of the test for differences in side effects between the two groups had a p-value of .13, Bradford is concerned that the value leaves a possibility that the drug does have side effects. He is unwilling, however, without more information, to second-guess the decision about the medication especially given his assumptions, 1) that the F.D.A. had approved the drug and 2) that the F.D.A. would not approve a drug with serious risk of side effects.

Bradford is hesitant to believe the findings of the ESP Study. He cannot find any flaws in the design, but he claims to have read studies with competing findings. He acknowledges that the difference of using dynamic pictures may result in the difference in findings between this study and others he has read. He specifically mentions that he would like to read critiques of this study before developing a complete opinion on it. After a discussion of the ESP replication studies, Bradford states that, "generally, studies that are well done and replicated and… don't have any serious critiques against them… would do a fairly good job of convincing" him of something. He then claims that the major problem he has in being convinced by the ESP Study is that he cannot think of a rational explanation for the results.

Bradford says the findings of the Lyon Diet Study are not surprising. He critiques the experiment for its sloppiness. He says that, if the researchers wanted to prove that the Mediterranean Diet is more advantageous than other diets, they should not have used a typical American Diet for the control group. They should compare it to a healthy diet as recommended by American health professionals to their patients. If they wanted to show

the health benefits of aspects of the Mediterranean Diet, they should have separated the individual factors, like the use of olive oil or the consumption of alcohol, for analysis, rather than looking at the Diet "holistically."

The Lyon Diet Study results seem broader in scope to Bradford than other articles about diet he has read; it attempts to make too many conclusions. He has not read diet articles in "random magazines," like Readers Digest. He prefers scientific articles to articles that are marketed to the general population. Magazines targeted to a general audience contain articles written to sell the magazines. Bradford believes that the best diet is a well-rounded diet, but that the general public wants to believe that there is a simple solution to diet issues so what is published in popular magazines are "magical solutions" to health problems that do not actually exist.

Bradford is consistent in saying that the p-value is the probability that the null hypothesis is true, given the data collected. He was able to choose the correct meaning of the p-value from a list on the final exam, but he is unable to produce it on his own two months later. Bradford was able to reason clearly about the p-value associated with the side effects. He realizes that one wants to minimize the Type II error (accepting the null hypothesis when it is not true – in this case, deciding that the drug and the placebo account for a similar number of side effects when in fact the drug produces more) and that accepting the null hypothesis with a p-value of .13 is, perhaps, not appropriate.

Bradford does not include any discussion of experimental design in his written response to the Email from Dad. When asked what the doctor meant by "randomized, double-blind." Bradford correctly defines both terms. When asked if those features are important, he responds that they are. Further, he explains that if the study were not blind, this would negate the study of the placebo effect and that "double-blind" reduces bias on the part of the experimenters. Finally, he explains that randomization prevents bias in the

make up of the experimental and control groups. Bradford's first response to the ESP Study is to comment that he cannot find any flaws with the design. The first questions that Bradford asks about the replication studies are about experimental design. He seems to be satisfied that the study is well designed, but he does attempt to comment on the use of a one-sided test rather than a two-sided test. In the end, he decides that he does not know enough about how to set up a statistical analysis and trusts that the researchers must have "done it the way they were supposed to." As far the experimental design of the Lyon Diet Study, Bradford seems to believe that the study was properly designed, but attempted to test too many factors in one study.

1.2 Kyle

Kyle is a 19-year-old freshman pre-pharmacy student of Asian descent. He has a 4.0 G.P.A, and reports that the A he earned in statistics is too high compared to what he learned. He had a lot of difficulty in discussing the statistical results, so this may be an accurate assessment. Kyle tested as a HighNC-Transitionalist in September but was a HighNC-Evaluativist in February. Kyle did not seem to connect well with the first two tasks. This may be because he is not close to his one living grandparent, his family does not discuss health issues in depth, and he seems unconcerned about his health or his future. He does not worry about what or how much he eats, although he does try to eat one good meal a day. His planning for the future is limited to thinking about the courses he needs to take to be ready to enter pharmacy school in a few years. When asked about his belief in ESP, he first asks for clarification of what ESP is. When it is explained, expresses skepticism about the existence of ESP, although he will not rule out the possibility that it does exist.

Kyle does not suggest a course of action for Grandma's medication in his email response, although he does write that the study shows that the medication is effective.

When probed in the interview, Kyle's first response is that the small p-value is the evidence for the drug's effectiveness. He also mentions the drop in blood pressure in terms of the raw measurements. When asked for his overall assessment of the medication for Grandma, Kyle asks whether the drug is experimental. In the case where he is presented with information that the drug is experimental, he worries that the one study might not provide enough information and is specifically worried about the existence lurking variables and the possibility that the results of the study are wrong. He says that he would be less worried if a few studies had been done. When asked how his response would be different if the F.D.A. had approved the drug, Kyle says it is probably a good idea for grandma to be on the medication particularly if she needs it badly. He does suggest that his father do research on alternative medications and look for studies that talk about side effects that might be detrimental to Grandma's condition.

In his writing, he attributes the stories of Larry and Sally to the drug not being perfect. When probed about the anecdotes in the interview, Kyle gives allergies as a possible reason for Larry's fever and states that Sally's reaction might be the result of outside factors, such as, stress or the onset of a cold. He suggests that his father might check into possible drug allergies as a result of this conversation.

Kyle's first comment about the Lyon Diet Study is that it is interesting but also harsh in its treatment of the death of subjects. He found it interesting to learn about a new diet and to think about the number of people that might die during a study. Because Kyle is currently healthy, it is difficult to get him to comment on the study results. When asked to imagine that he is older and not in good health, he says that, in that case, he would probably look into the Mediterranean diet. The number of deaths and number of incidents of cancer are the features of the data he claims are convincing evidence. He moves from that statement to a discussion of types of foods that are bad. For example, he

says that the Atkins diet would be bad for heart patients because it contains mostly red meat which, when cooked, usually has blackened parts that increase the risk of cancer and that the meat itself has a high fat content which is also not healthy. He says he learned about nitrates in the blackened parts that can cause cancer in DNA through lectures in a genetics class.

Kyle's reaction to the ESP Study is that it is "weird" and interesting and that he didn't expect the percentages to be so high. In particular, he thought it was interesting that a study would last for six and a half years. He asks a follow up question about the process and what the researchers mean by a "success" and then says that he has no idea what to think about the study. He cannot find a problem with the study, but he is unwilling to believe in ESP. He explicitly expresses the opinion that he is being hypocritical because he believes the results of the other two studies but is unwilling to believe the results of the ESP Study. He talks about the fact that one can tweak the results, for example, changing "greater than" to "greater than or equal to" without really changing anything, but says they have not done that in the ESP Study. He then says that he has not experienced ESP and does not have empirical data, numbers about ESP, it is just categorical data, yes-no. When he reconsiders the p-value, he seems torn about his opinion. Then he focuses on the number of subjects as a possible reason not to believe the outcome of the ESP Study. He cites a problem from his statistics text in which the number of subjects can be used to manipulate the p-value. The problem asked the student to change to the number of subjects and redo the problem and that drastically changed the p-value. Kyle claims that it is difficult to believe in ESP since he has never seen it and does not know much about it. He does not find the replication studies convincing. He would only believe in ESP if someone could read his mind or if a subject was correct on 90% of his trials. In the end, Kyle still feels as though he is being hypocritical by not

believing the results of the ESP Study, but he says that it is okay to be stubborn sometimes and that the pharmaceutical companies have "delivered" and he has experienced the value of medication first hand.

Kyle was unable to complete a sentence about a p-value, although he says it's the chance that something is either true or not true. Perhaps this is not surprising since on his final he reported that the p-value is both $P(H_0|D)$ and $P(D|H_0)$. During the conversation about the Email from Dad, Kyle says, without prompting, that the p-value of .001 shows that the results are "pretty dang accurate." When probed about the meaning of the p-value, Kyle says that "they" had to assume that something was incorrect and the p-value is the probability that it is correct. He follows this by saying that the researchers would have started with a well thought up assumption, which they would have used in the calculations. So, the p-value is the probability that they were actually wrong from the beginning. When asked what the assumption might have been, Kyle is unsure but says he thinks they may have assumed a value for the average drop in blood pressure without medication. His conversation with regard to the Lyon Diet Study is largely the same as that about the medication study.

Kyle does not discuss experimental design spontaneously. Even when the results of the ESP Study appear counterintuitive, he does not try to find problems with the experiment to argue against the study. In the middle of the discussion of the ESP Study, Kyle is asked to talk about "randomized" and "double-blind" from the Email from Dad task. He correctly defines the terms two terms. When asked why this is important he says that double blind is important so that the subjects will not react based on the knowledge of what they are taking so there would be less chance of error. He says that randomized is important because if you were to put all healthy people in one group and all sick people

in another that would be a terrible design. When we return to the discussion of the ESP Study, he does not transfer the discussion of experimental design to the ESP Study.

2. THE HIGHNC-TRANSITIONALIST: HANNAH

Hannah is a 21-year-old White senior, majoring in advertising and English. She has a 3.89 G.P.A. and reports that the A she earned in statistics is too high compared to what she feels she learned. Hannah is candid about the fact that she did not attend class, did not take the second midterm and studied for the final exam three days prior to taking it. She feels that statistics, as a subject, is important and regrets not having had enough time to learn it properly, but is not committed to doing so in the future. In September, Hannah tested as a HighNC-Multiplist; in January, she retested as a HighNC-Evaluativist. Hannah's outlook in January may have been affected by a course she took in lying and perceptions in which they discussed the nature of knowledge and facts. Knowledge seems much more of an illusion to her now than she remembers it being prior to the class.

Hannah has a stake in two of the three tasks, the Email from Dad and the ESP Study. She distrusts doctors because of personal medical problems that she believes she solved with the help of the Internet, rather than through the help of the doctors. While her Dad is not likely to ask questions like those posed in the email, her maternal grandfather, who had diabetes and high blood pressure, lived with the family when he was dying, and there was much discussion of his care in the house. Based on family experiences, Hannah has a prior belief that ESP exists. Hannah seems less connected to the subject of the Lyon Diet Study. While she has a general sense that one "should" eat a healthy diet and what defines a healthy diet, she is not particularly concerned about her health or eating habits. Also, at this time, she does not plan very far into the future, being

more concerned with near term decisions, like choosing a major and completing class assignments.

In her written response, Hannah advises her father to have Grandma try the medication based on the good success rate. When asked what she meant when she wrote that it sounded like the drug had a very good success rate, Hannah responds by saying that the drug lowered blood pressure by almost twice as much as the placebo. This seems, to her, like a pretty significant effect. In her writing, Hannah does not seem particularly concerned by the anecdotes of Larry and Sally. Hannah tells her father that if Grandma does develop side effects or if her blood pressure does not fall, she can be taken off the medication. She also advises that if the anecdotal evidence has caused her father or her grandmother stress, he might want to investigate other medications with which he might be more comfortable because the stress would not be good for their blood pressure. The existence of Larry and his fever prompts Hannah to mention that side effects could, potentially, be an issue. When Hannah is told of the non-significant results about side effects, she suggests that perhaps Larry had some other condition or a drug interaction that caused his fever. Hannah believes that Sally is an example of a rare case. When probed, Hannah explains that it is unlikely that the drop in blood pressure as large as was found in the sample would have happened if a lot of subjects actually experienced a rise in blood pressure. Therefore, Sally must be an unusual case.

Hannah finds the results of the Lyon Diet Study "shocking" and "pretty extreme." She says it shows that the American diet is contributing to a lot of the problems we are facing in this country. She has a notion that the typical U.S. diet is not optimal in terms of health, but the size of the benefit found surprises her. When she is probed to explain what information was compelling, Hannah first discusses the difference in the number of deaths and occurrences of cancer. She goes on to say that the small p-values show that

the results are pretty accurate. While she is pleased to see the results, she is interested in knowing about other factors, such as how people felt and whether they had more energy while on the Mediterranean diet.

Hannah finds a large difference between the average diet articles published in popular magazines and the Lyon Diet Study article. She says that the magazine articles are "based on absolutely nothingness" and that it is important to have all of the information of the kind provided by the Lyon Diet Study article. She believes it is a question about how much effort one wants to put into one's beliefs. She thinks that people are trained believe whatever they hear and trust the sources, but that people should probably be making more critical decisions incorporating the reliability of the sources and looking at the data that is provided.

Hannah is also surprised by the results of the ESP Study. She is reluctant to believe the results with regard to static images because she cannot think of a reason that static pictures and dynamic pictures would produce different results. She also believes that anecdotal evidence of ESP is more convincing evidence of its existence than is a scientific study. When replication studies showing no significant results are presented, she does not question her belief in the existence of ESP. She says that ESP is not a phenomenon that can be scientifically studied, or at least not using the technique described. At the beginning of the conversation about the ESP Study, Hannah asks questions about the procedure. A summary of the study, by her, to an imaginary friend, indicates that she understands the procedure and rationale behind it. Hannah uses meta-cognitive skills to diagnose her own reaction to the tasks explaining that she believes the results of the ESP Study because they matched what she wanted to believe.

Hannah has a very difficult time discussing the p-values. She knows that when they are small, something unusual has occurred. She also associates a small p-value to

"people" in the tail of a normal distribution. She seems to believe that a small p-value tells you something about individuals in the population, but also talks about the p-value as the chance that an error has been made. In the Email from Dad task, Hannah's initial response is that the drug showed a good response rate. When asked what she meant by that, she responded that the drug had a significant effect based on the difference in blood pressure drop between the control and experimental groups. When she is told that the p-value associated with the difference in side effects between the two groups is .13, she tells her father that that is a good results and means that there was no difference in side effects so the medication does not produce more side effects than the placebo.

In general, Hannah does not spontaneously discuss experimental design. When prompted in the Email from Dad task, Hannah correctly defines double blind. Her definition for "randomized" is actually a definition of choosing a random sample. Without prompting, she explains that choosing a random sample would ensure that the study results were not biased to one particular ethnicity, for example. When prompted to discuss the value of a double-blind design, Hannah reports that it would reduce bias. When asked about the design of the Lyon Diet Study, Hannah pauses for nearly half a minute. After the question was rephrased to have her compare the interview article to other diet articles she has read, Hannah mentions that the article says the phrase "prudent diet," describing the diet suggested to the control group, was not well defined, but that she does not have a problem with the fact that the study modified the Mediterranean Diet, although perhaps it indicates that the true Mediterranean diet is too difficult to follow. She seems to understand the design of the ESP Study, but suggests that ESP cannot be studied scientifically. As part of the discussion of issues with the design, she mentions randomization. The randomization of subjects, she suggests, may dilute the effects of ESP because it may be the case that only some people have the "gift."

3. THE HIGHNC-MULTIPLISTS: SARAH AND MEGAN

3.1 Sarah

Sarah is a 20-year-old White junior pre-pharmacy student with a 4.0 G.P.A. She reports that the A she earned in statistics accurately reflects her learning, but late in the interview she makes a remark about her statistical knowledge being nonexistent. It is unclear whether the statement is an act of modesty or whether it occurred as a result of difficulties she encountered during the interview or neither. Sarah tested as a HighNC-Evaluativist in September and retested as a HighNC-Multiplist in February.

As Sarah works at the Heart hospital, she has relevant background knowledge for the Email from Dad task. While it is not the type of discussion she would have had with her family in the past, they might ask her now that she has some experience in the field. She is close to her one living grandparent. In discussing diet, Sarah reports that she would like to eat in a healthier manner than the time and budget constraints of her student life allow. She reads a lot of diet articles because she wants to be healthy and not die at an early age. Of all the subjects interviewed, Sarah seems to have thought the most about the future. She is decided on both career and partner. She knows where she wants to live and about the type of life she would like to have. Sarah describes herself as a logical person who needs facts to change her opinion. Logic made her disinclined to believe in ESP.

In Sarah's written response to her father, she advises her father to talk to the doctor about other possible medications that do not have serious side effects. She says that the results of the study show that the drug has a "really high chance" of lowering blood pressure. She suggests, however, that her father ask the doctor how many subjects experienced a rise, instead of drop, in blood pressure. When probed, Sarah seems to be less concerned about Sally, saying that if grandma's blood pressure rose, they would simply look for a new medication. The fever that Larry experienced was much more

alarming to Sarah. She expresses a concern that at Grandma's age, she might not survive a fever high enough to require hospitalization. In the end, she advises her father to keep grandma on the medication, but to watch for side effects and do research on the side effects of this medicine and on other possible medication that work well and do not have side effects.

Sarah's general notion of what is healthy seems to have as its evidentiary basis factors that are "common knowledge" and have stood the test of time. The result of the Lyon Diet Study confirms some of her previous views about diet, that the American diet is bad because it contains too much red meat and fried foods and not enough fruits, vegetables, fish and good oils, like olive oil. Sarah claims that, for her, the data, like the kind provided in the article, are more compelling than an anecdote someone had written on the internet, but that the American public would not know what the numbers mean. She says that the use of the statistics and the language that is used in writing statistics might seem to the general public like jargon or as if the researchers are trying to trick the readers. She admits that she learned in her statistics class that one had to be very careful when reading statistics because there are a lot of ways to "twist" the numbers and make the results look like "something they are not."

Sarah's reaction to the ESP Study is that it is "weird." She wonders how the researchers came up with the process used and is surprised that the results for dynamic targets are better than those for static targets. It seems to her that focusing on only one thing would be easier. Sarah says that the 37% success rate, especially given the large number of trials, is very significant, but then she says that she is just too logical to believe in ESP. She claims the study might have been completed on a "good guessing day" and then says that she might believe in ESP if the success rate were 75%. She also brings up the notion that the study might be biased if some of the targets played into people's

subconscious. After a lengthy discussion of two replication studies, in which she is only given the p-value and number of trials, and from which she is able to accurately infer the percent of successes, Sarah seems willing to entertain the possibility that ESP exists.

Sarah is unable to discuss the meaning of the p-value in context, although she did say that a small p-value proves the alternate hypothesis while a large p-value proves the null hypothesis. She has a very strong notion of p-value as significance level. When told that the difference in side effects was not different between the control and experiment groups in the Makemewell Study with $p = .13$, Sarah says that that value is a little high for a p-value especially given how important side effects would be to grandma. When she is reminded that the results were not significant, Sarah counters by saying that researchers could pick any value against which to test significance and find no significance, but the results would still be significant to the actual people and families of the people taking the medication. Sarah is very adept at navigating between the raw data and p-values, showing a lot of conceptual understanding of the sampling distribution. She is interested in knowing both the raw data and the p-values as part of the decision-making process because she has learned that one can present the statistics in a misleading manner if the data are not in evidence.

In the first two tasks, Sarah does not mention experimental design without prompting. When asked in the discussion of the Email from Dad task, she defines "double-blind" and discusses the placebo effect and the reduction of bias based on the doctors not knowing what medication each patient was receiving. She does not attempt to define randomization, although that was specifically included in the question posed by the interviewer. When asked what was well done or not well done about the Lyon Diet Study, Sarah makes only the general comment that it seemed pretty thorough and then proceeds to discuss the results with regard to the diet's effect on cancer and heart health.

Later in the interview, however, when asked why the statistics appeal to her more than anecdotes, Sarah says that with an anecdote, one cannot know if the diet will work for people with other body types, but the fact that the subjects in the Lyon Diet Study were random people made the conclusions more generalizable.

When discussing the ESP Study, Sarah alternately attends to and ignores elements of experimental design. In the beginning of the conversation, she asks many questions about the participants, whether they were novices or experts and whether they were randomly assigned to be senders and receivers. Later however, she talks about the possibility of selection bias, without having realized that the computer had chosen at random the targets and the choices for the receiver. Both the questions and the later misunderstanding might be attributable to lack of careful reading.

3.2 Megan

Megan is a 19-year-old White junior double majoring in public relations and government. She has a 3.97 G.P.A and reports that the A she earned in statistics accurately reflects her learning. She tested as a HighNC-Transitionalist in September and as a HighNC-Multiplist in February. She was raised in a conservative, religious home and admits that college has given her a chance to think about and change her beliefs. She claims that her family was instrumental in shaping her beliefs, from those about ESP and health to those about abortion and capital punishment. She has retained the beliefs she developed from her family about eating healthy and exercise as well as capital punishment, but she has already changed her opinions about abortion. She took a class during the Fall semester called Contemporary Moral Problems in which the students read two conservative opinions and two liberal opinions on the same issues. Those readings helped her to evaluate her beliefs. She claims that she does pay attention to author and

strength of argument when evaluating new ideas. Given all of this, her NC classification and the changing nature of her EU classification seem reasonable. She is interested in formulating opinions based on facts and she is in a very transitional phase of her life as she tries to separate her own beliefs from those with which she was raised.

None of Megan's grandparents are alive and she was not close to them when they were. The Email from Dad task was quite abstract, as her family does not discuss health concerns and her father would know how to find the answer to his questions himself. Megan thinks that doctors are smart and do a reasonable job, though she does not like going to them because she believes that they do things to her that are unnecessary. Megan gives a detailed account of her diet. She claims to be a "health nut;" she exercises and thinks a lot about what she eats so she does not end up eating junk food all the time. She will not eat at McDonalds because of the book and movie, Fast Food Nation. She is worried about getting Alzheimer's because of a movie she saw, but other than that says she does not think about the future. Megan calls herself "narrow-minded" when she says that she has never experienced ESP and does not believe in it.

Megan advises her father to keep Grandma on the medication and correctly interprets the results of the study, including a description in context of the hypotheses tested and the meaning of the results. She suggests in her writing that the experiences of Larry and Sally may have been the result of other factors and that, while all drugs have side effects, if Grandma has been on the medication without any adverse reaction, it is probably safe to keep her on the medication. She specifically mentions in the interview that the results of the study are very strong, based on what the doctor said about the significance level and the effect the medication had when compared to a placebo. She also says that she trusts the doctor much more than the next-door neighbors.

Megan "likes" the Mediterranean diet described in the Lyon Diet Study because it includes foods that she likes, such as fruits and bread. She thinks that a diet that cuts out food groups cannot be healthy. She admits that she does not read diet articles so all diets could "sound good" to her, but she is impressed and convinced of the value of the diet on cancer and heart health by the results given in the Lyon Diet Study article. Her ideas about healthy diets are based on what she learned from her parents. When probed about how she might discuss the Lyon Diet Study results with her roommate or an old boss, Megan indicates that she would mention the health benefits with the roommate, but not with the boss. Megan says that encouraging him to eat a healthy diet would be a behavior at the edge of what she would be comfortable doing; discussing the lower risk of death, she says, might make her seem "psycho."

Megan's first comment about the ESP Study is that, while she is not a "big ESP person" the results of the study indicate that perhaps ESP does exist. When asked how she came to that conclusion she cites as evidence the significant findings on the differences in results between static pictures and dynamic pictures. This significant difference indicates to her that something must be happening. In her final consideration, she maintains that she is still not an "ESP person" and would like to see results of more studies, but the results presented have opened her mind to the possible existence of ESP. When the replication studies are presented, she is impressed. She then says, however, that she would really want to experience ESP in order to really believe it. When asked how many times a person who claimed to have ESP would have to choose the correct target, Megan says that twice would be sufficient. Throughout the conversation, the existence or on existence of ESP appears to be not particularly important to Megan; she does not seem to be particularly invested in the topic.

Megan's description of the p-value is correct and consistent all the way through the written task and the interview. When discussing the Lyon Diet Study, she says that the numbers that are presented are confusing because there are so many of them, and that p-values are both more understandable and compelling.

Megan defines "double-blind" in her written response to the Email from Dad but neither discusses "randomized" nor explains the value of a double-blind design. When probed in the interview, Megan says that a double-blind design prohibits the doctors from skewing the results and insures that the difference in effect found by the drug is really attributable either to the drug or psychological factors. When asked to define "randomized," Megan has a bit of trouble with her response, but then gives an essentially correct response. Similarly, when probed about the value of a randomized design, Megan struggles initially and then explains that it allows the results to be more generalizable. When discussing the Lyon Diet Study, Megan does not mention experimental design. Megan does not spontaneously discuss experimental design features of the ESP Study. When probed about the design of the study, Megan says that having more target choices, she suggests 30, would make the study more compelling. She also mentions that asking people questions, such as rating the targets, is less reliable than systematic measurement that is done in drug testing, for example.

4. THE MIDNC-EVALUATIVIST: JEWEL

Jewel is a 21-year-old White senior majoring in psychology. She has a 3.9 G.P.A and reported that the A she earned in statistics accurately reflects her learning. She tested as a HighNC-Multiplist in September and retested as MidNC with an unusual EU grouping, EEAA, in February. The fact that all of her choices either exhibited an Absolutist or an Evaluativist perspective seems to indicate that she does believe that assertions can be judged either against truth, in the case of absolutism, or against each

other, in the case of evaluativism. According to the classification, statements about the physical and social worlds can be compared to the truth. Aesthetic and value judgments can only be compared to each other. At the end of her interview, Jewel says that understanding statistics is "ridiculously" important in general daily life because, otherwise, anybody who "throws numbers out there at you, can convince you of anything."

Jewel thinks that, aside from some "quacks," doctors are generally good people. She does have living grandparents and talks to her mom about their conditions and medicine, but the questions have never been about specific statistical results. She is a vegetarian. When she became a vegetarian she saw a nutritionalist to obtain information to assure herself and her family that she was not "going to die" from not eating meat. Jewel does not believe in ESP, although she is not particularly emphatic about her disbelief. She reads her horoscope, but only for fun, not expecting the things it predicts to happen. Her planning for the future is largely abstract. She thinks in general terms about decisions she will face and changes that will occur but she does not have a set plan, other than a desire to enter the Peace Corp and then, perhaps, to go to graduate school.

Jewel's initial reaction to the Email from Dad task is to write to her father that she does not trust her understanding of the information, but would be happy to talk to her statistics professor to get some advice. When prompted to talk as best she can about the situation, Jewel points out that 40 subjects is a rather small sample on which to be making decisions. When asked to explain the p-value found by the study, she claims that the study shows, fairly conclusively, that the lowering of blood pressure on the part of the subjects was due to the medication and not chance. When queried again about the sample size, Jewel says that the results would be more conclusive if the same p-value had been found for a larger sample size, such as 400.

Jewel tells her father that Sally's reaction to the medication is variation about the mean and nothing about which to be concerned. The statistics, she says, are based on averages, but individual people will be both above and below the mean. Jewel says about Larry, that he was experiencing side effects and that the doctor should have told her father about possible side effects. On being informed that a test for differences in the incidence of side effects between the two groups was not significant with $p = .13$, she explains that this means that the medication did not produce a higher percentage of side effects than did the placebo.

Jewel is basically happy about the Makemewell for Grandma, although she would like to check the calculations of the p-value reported by the drug company. She reiterates that Larry and Sally are just two of many people who take the medication and that her father should not be too concerned about them. If he is, she suggests that he do more research, online and in the medical journals, about the medication.

Jewel's first reaction to the ESP Study is to question whether it is a real study. She seems not to have completely understood the article and asks questions about the design. She finds it strange that someone would study a phenomenon like ESP scientifically. Although Jewel tries to make an argument against the study results based on the fact that only 4 choices given, she concedes that the study is "statistically valid" and that the authors specifically mention the 25% chance of guessing correctly. She credits the authors with having been published in a well known, peer reviewed journal, admitting that this lends credence to their conclusions. Given those points in favor of the research and after admitting that the results for dynamic images were highly significant, Jewel says that she would want to hear about replications in order to be convinced of the existence of ESP. She is not convinced, however, by the existence replication studies. She says that she would need to watch a study or experience the phenomenon herself,

although she freely admits that that would be less valid than a study. In the end, she thinks the possibility of ESP is "cool," but is not convinced that it does exist by the evidence presented. She cautions that even studies published in journals by well respected researchers have been found to be false or incorrect at some time in the future.

Jewel finds it interesting or unusual that people have studied, scientifically, both ESP and diets. She finds fault in the fact that the Lyon Diet Study assesses a general diet plan and not specific changes in diet, like reducing red meat intake or increasing fiber, but she is impressed with the thoroughness of the study and the number of subjects that were tested. She thinks it is "funny" that someone who must like Mediterranean food went out and tried to prove that it is good for you and is not impressed that the findings show that the diet is better than a typical American diet. She is more convinced by her previous research on diet, which was done on the Internet, then she was by the Lyon Diet Study. She admits that none of the information she gathered was the result of a scientific study, but is convinced of its validity by the number of sources and the general level of acceptance that certain ideas have gained in the public eye.

In general, Jewel does not bring up the p-value on her own except to cite that the result was significant. She only discusses its value when prompted in the Email from Dad task. She has a lot of trouble talking about the meaning of the value and then eventually says that there is a 99.9% chance that the results are due to the drug and not chance. She does not mention the p-value when talking about the other articles and was not prompted to do so by the interviewer.

When asked about the experimental design features of the Email from Dad task, Jewel is able to define both "randomized" and "double-blind" and she spontaneously explains their value in providing validity to the study results. She raises the question of the use of only four targets in the ESP Study, but does not follow through on that line of

reasoning at any point when she is trying to make sense of the results. She does not discuss any of the other design features of the ESP Study, nor does she discuss the design features of the Lyon Diet Study except to mention the value of the number of subjects who were studied.

5. THE MIDNC-TRANSITIONALIST: MARIPOSA

Mariposa is an 18-year-old White sophomore majoring in nursing and nutrition. She has a 4.0 GPA and reports that the A she earned in statistics accurately reflects her learning. In September she tested as a LowNC-Multiplist. In February she tested as a MidNC-Transitionalist. The change in EU classification is in the expected direction, developmentally. The difference between September and February is in the category, "truth about the physical world," in which Mariposa transitioned from an Absolutist stance to an Evaluativist stance. Her outlooks in the other domains did not change. The change in NC score is just over one standard deviation.

As a nursing student, Mariposa has a favorable opinion of the medical profession. While she is currently taking a nursing research class, she has not yet read many pharmaceutical trials. Her Dad is unlikely to ask her about mathematics, because he and her brother are engineers. He might ask about medical issues because she is studying nursing. Both of her grandmothers are still alive and have no health issues. Both of her grandfathers did have health issues. Mariposa worries about her health and diet because her genetics are "bad" and her nutrition classes have scared her into eating well. Her parents have high cholesterol. She thinks about the future in terms of making decisions now based on what will be best for her future. She is certain that there is no such thing as ESP.

Mariposa tells her father that the medication is reasonable based on the evidence from the study, which she says is quite significant. She seems to have difficulty with the

concept of causation; in the same sentence she says both that one cannot infer that the medication caused the drop in blood pressure and then that it is the major cause. In her letter, Mariposa tells her father that Larry and Sally may have other medical conditions that caused the side effects. When questioned about this she cannot suggest a way to find out if Grandma is susceptible to the same effects. She reiterates that the effects on Larry and Sally might have been the result of an interaction with another factor or medication, but that the results of the study are reliable and Grandma should stay on the medication unless she developed side effects or her blood pressure began to rise.

Mariposa finds the Lyon Diet Study "interesting." She is surprised by the number of people who died during the study and by the difference in cholesterol. She said that she would use the article if she had a conversation with her family about changing their diet to a healthier one. Mariposa said it is the p-values that are the compelling evidence that the diet is effective. When comparing the Lyon Diet Study article to the types of diet segments she had seen on TV recently, she says that the Lyon Diet Study article is more compelling since it provides proof and justification for its value in the form of an experiment. When someone on television says that a diet worked for her, she might not be telling the whole truth, and her results might not work for everyone.

Mariposa finds the ESP Study difficult to understand. She feels as though the researchers have "thrown a lot of stuff" at the reader to confuse her. She is sure it was a good experiment because they obtained the results, but she is not convinced of the existence of ESP. After a discussion in which Mariposa expresses concern about the design of the study and those questions are answered by the interviewer, Mariposa admits that it sounds as though the researchers were trying to "get rid of bias," but she wants to hear about some follow-up studies before she will believe the stated results. When she is presented with replication results in which the p-values were higher, she finds those

results more believable because the chance of the results occurring without the existence of ESP is in a more reasonable range. She uses the words "believable" and "unbelievable" to describe the results of the studies. The ESP Study results are unbelievable because they showed a .02% probability which is certainly too low for ESP. The results of the imaginary replication are more believable because they showed a 2% or 3% probability. She is still not compelled to believe in ESP. In the end, she says that she is more compelled by the Lyon Diet Study than the ESP Study, although she has similar problems with the experimental design of the two studies, because the Lyon Diet Study results match information she had heard previously, but she has never been exposed to information confirming the existence of ESP.

Mariposa has trouble talking specifically about the meaning of the p-value. She describes it as the probability that the results are "outside of chance" and says that it represents the probability that the results were due to chance and not the experiment. She starts to claim that it is the complement of the probability that the alternative hypothesis is true, but realizes that this interpretation is incorrect. She does have a sense that small p-values are significant and that, in a well designed experiment, a small p-value means that the results have not occurred by chance.

Mariposa correctly defines both "randomized" and "double-blind" in her written response to Dad. She writes that randomization would produce less of a bias and defines bias as researcher influence. She does not explain the value of the "double-blind" design. Mariposa mentions the experimental design factors as the first part of her response to the question, "what evidence is there that Makemewell is effective?" When probed, she reiterates that the experiment design factors make it more difficult for the researchers to present untruthful results. She does not mention experimental design spontaneously when discussing the Lyon Diet Study. When probed, she responds that the Lyon Diet

Study is randomized but clearly could not be double blind. She criticizes the Lyon Diet Study because it did not have enough control, but still finds the results compelling. When she is trying to create an argument against the ESP Study results, she questions whether the study was randomized or double blind, but that line of inquiry is not sustained.

6. THE MIDNC-MULTIPLIST: NATALIE

Natalie is a 19-year-old sophomore pre-pharmacy student of Indian decent. She has a 3.64 G.P.A. and reports that the A she earned in statistics accurately reflects her learning. In September she tested as a HighNC-Multiplist and in February she retested as a MidNC-Multiplist. The drop in NC score was almost two standard deviations; her score was just above the upper quartile in September and below the median in February. The conversation with Natalie is noteworthy for the way in which she responds to most of my questions with a clarifying question or by simply repeating my question as if making sure she knows what I had asked. She appears to spend the bulk of the interview trying to give the answers she thinks I want. She specifically asks, while completing the EU measurement task, if there are "right and wrong" answers. We also spend a lot of time clarifying the studies. She seems to have difficulty reading critically and formulating opinions about the information. She expresses no strong opinions. This is, perhaps, cultural or due to the way she was brought up.

Natalie lives with her grandmother who has high blood pressure so the first task should have been quite relevant. Natalie's grandmother discusses her aches and pains and medication with her. Out of curiosity and not for a class, Natalie has read a few pharmaceutical trials. Natalie says she "kind of, sometimes" believes in ESP and that her sister says she has it but is just joking. Natalie describes her diet in detail, but says she is not that careful about what she eats. She worries about aches due to physical activity,

which may prompt her to drink more milk or take vitamins. Natalie's planning for the future is quite general; she thinks about getting into pharmacy school and whether she will need a job to provide for herself and her family. Natalie claims to have heard about the Lyon Diet Study previously. In particular, she says that told her grandmother about the Mediterranean diet and her grandmother tried it but gave it up after a week.

In her writing, Natalie advises her father that Grandma should stay on the medication "because the doctor's words and statistic are more valuable and accurate than the neighbor's personal statement." When probed about this, Natalie says that the neighbor is only one person, who may have been affected by environmental factors, but the doctor has worked with many subjects for a long time. She both writes and says that if Grandma's blood pressure rises, she should return to the doctor and tell him of her conditions.

After she finishes reading the ESP Study, I explain the experimental design and the results to her. Once she claims to understand the study, Natalie's first comment is that there are not enough subjects. Next she says that 37% could still happen by chance, and that 45 or 50% would be better. When asked for clarification, Natalie agrees that a 45 or 50% success rate would be high enough for her to begin to believe in ESP. After she discusses the p-values associated with the study, she says that they indicate that it might be a "good study," but that they still need about 200 more people. When she is told about the replication studies, Natalie asks more questions about the procedure used in the study. At the end of the conversation, Natalie says that there may be ESP for some things, but not for everything. Then she apologizes for the conversation having taken so long.

Natalie's first comment about the Lyon Diet Study is to ask who the subjects are and whether they are all Americans. We discuss the procedure and results of the study. Then as a response to the question, "what do you think about the article?", Natalie

responds that she thinks that if a person has a heart attack, he should watch his diet more and that doctors should give their patients guidelines about what to eat. When asked about the type of guidelines that might be given, Natalie mentions that they should be told how much fish and protein they should eat in a week because meat is not good for them. When asked how she knows this, she says she learned it from her grandmother who is scared of having a heart attack. She says that she had heard about the Mediterranean diet on television, where they gave counts without p-values. For her, she says, the p-values are more convincing. When probed she says that the television reports "for everybody" and "not a lot of people know" about p-values or can compare them.

Natalie's discussions of p-values are not consistent. In the Email from Dad task, she tries very hard to express the meaning without being able to do so and finally asks to skip the question. The only coherent utterance is something like ".1% chance of happening" which she says twice in her attempt at a response. She correctly identifies the meaning of the p-value in the ESP task, the probability that the results would have occurred without ESP. In the Lyon Diet Study, I ask her not to attempt to give the answer her statistics instructor would be expecting. She responds that the small p-value proves that a Mediterranean diet is better. Note that she did respond to the multiple-choice exam question about the meaning of p-value correctly.

It is interesting to note that Natalie is not at all concerned with the 80 participants in the drug trials, but when we discussed the ESP Study, she says that 160 subjects is not enough to be convincing to her. The only other time that she spontaneously mentioned experimental design is in asking about the heritage of the Lyon Diet Study participants. She does not write about the experimental design features in the Email from Dad task. When asked was "randomized, double-blind" means, Natalie begins to answer haltingly, then remembers I am supposed to be her grandmother. She then defines a blind study

and mentions a placebo. When she attempts to say why it is important, she is unable to say anything coherent. I spend much more time in this interview discussing the procedures used and the results than in any of the other interviews.

7. THE LOWNC-EVALUATIVIST: CHARLOTTE

Charlotte is a 20-year-old, Chinese American junior majoring in Chinese and Spanish who took statistics to fulfill the requirements of the Bridging Disciplines Program in public policy. She has a 3.72 GPA and reports that the C she earned in statistics was too low compared to what she felt she learned. Charlotte tested as a LowNC-Evaluativist in both September and February. Charlotte professes that she hated her statistics class. She knows she could have done better in the class because she "pretty much understood the majority of the information." She says that it was "just hard to understand" on her own since she had never taken a statistics class before and after the first half of the semester her boyfriend could not help her anymore and she had to teach herself the rest of the material.

All of the tasks are abstract for Charlotte. Her father would not talk to her about her grandparents' conditions. He tends to handle those things himself and he would probably understand the doctor himself. Her grandparents have medical problems but she has not seen them in six years. She thinks doctors are "pretty well informed," but sometimes jump to conclusions about certain products. She is not concerned about her diet, except that her boyfriend eats a lot so she worries that she might gain weight from eating with him. Charlotte is a bit worried about the fact that her life in Austin is more sedentary than her life at home. She thinks about the future in general terms, having a job and kids, being financially secure, and saving for retirement. She believes that people have the capacity for ESP, but she is not sure that there is a way to test it scientifically.

In her response to the Email from Dad, Charlotte writes that Grandma should stay on the medication even though he knows two people who have suffered side effects. She cites the results of the study as evidence. When probed about this, Charlotte says that the drug is reasonable based on the "statistics." When asked what the numbers mean, Charlotte explains that there were 40 people, and that the researchers were comparing the difference between the drops in blood pressures of the two groups. She says that if you compare the two numbers, there is a significant difference. When probed about the anecdotes, she cautions that these are only two people out of hundreds who probably have not had side effects. She says that he should not let it "blow up" in his mind just because it is people surrounding him. She likens this reaction to the news of an airplane crashing because you hear a lot about it when it happens, but it actually does not happen that much. This leads to a discussion about the incidence of side effects. When Charlotte is presented with the side effect study results, she discusses the use of significance level in hypothesis testing. In the end she says that "if it was a .05 alpha level, then it would not be significant," but she thinks that is okay. At the end of the conversation, Charlotte reiterates that grandma should stay on the medication and, if side effects do develop, they could take her off and find another medication.

Charlotte's first reaction to the ESP article is that it is "boring" because she would rather read stories than numbers or technical things. She asks if I want her to talk about the p-values. When asked for her overall impression of the study, she says that she does not know if you can actually prove ESP. Then she says the study is okay, but might be biased. We discuss the procedure and then Charlotte says that the use of novice subjects might introduce bias, but she cannot explain how or why. While Charlotte will not rule out the possibility of the existence of ESP, this study does not convince her of its existence. She thinks the p-values should be smaller because these results could have

happened from random guessing. When asked what would convince her of the existence of ESP, she says that she would want to see a trial of a sender and a receiver, both isolated from each other and the audience, in which the sender looks at something and the receiver "gets it exactly." This would need to be done 10 times, all with correct outcomes, otherwise it might be "a chance of luck."

Charlotte's first response to the Lyon Diet Study is to say that she thinks that her mother was on a similar diet once. It was not for heart reasons, but to lose weight. Charlotte says her mother made up the diet herself, eating very little meat, not many carbohydrates, and a lot of vegetables and fruit. When asked about the results of the study, Charlotte says that the first thing she noticed was that it helps prevent the development of cancer. In particular, it stuck out that of the 14 cancer occurrences, only 2 were from the experimental group. Charlotte took a nutrition class last semester in which they discussed fad diets that could be "debunked" and those that had value. They used as evidence in that class both biological explanations and advertisements. If an advertisement made "claims" that were "way out to there," then they are probably not true.

It is very difficult for Charlotte to discuss the p-value, particularly in context. In her written response, Charlotte writes that there is only "a .1% chance of error" or a ".1% chance that the drug does not work. In the interview she says that "you can reject the opposite results by a .001 chance, but then amends herself to say that the smaller the p-value the more likely it is that the drug works. When given the results of the side effects study ($p = .13$, not significant), she says that one must chose an alpha level in order to decide whether a p-value is significant and then guesses that the researchers must have been using a 0.1 alpha level. In response to what that told her about Larry or Sally, she says that it means they are among the 13% who experience side effects. When asked

about the p-values in the Lyon Diet Study, Charlotte says the p-value is the likelihood to reject the notion that the diet helps prevent cancer. She goes on to say that there is "a 5% chance that you should reject it." After a lengthy conversation, agrees that what she means is that there is a 5% chance that we should reject that the diet works.

Charlotte does not write about either randomization or double blind design in her email. When asked about them in the interview, she says that they are "really good" and defines them both correctly. When asked to explain why they are good, Charlotte explains that a double blind design keeps the researchers from introducing bias when they hand out the medication. She says that the ESP Study is "okay," but the use of both novice and expert subjects might introduce bias. She is able to describe an experiment to test ESP and give the significance level that would convince her of the existence of ESP. Charlotte does not discuss experimental design when talking about the Lyon Diet Study.

8. THE LOWNC-MULTIPLIST: KEFIRA

Kefira is a 21-year-old White senior majoring in advertising. She has a 2.89 GPA and reports that the A she earned in statistics accurately reflects her learning. In September she tested as a LowNC-Transitionalist. In February, she tested as a LowNC-Multiplist. Kefira has the lowest GPA, by far, of any of the subjects in the Targeted Study. Often during the interview, she talka out loud to herself, either rereading the article, or verbalizing what she is trying to remember having learned in the statistics class.

Kefira does not seem to have strong opinions about any of the task topics. Her grandparents have been dead since she was about three and her mother would not leave the doctor's office without understanding the information she is told. Kefira does not pay much attention to her diet and does not think very much about the future except in general terms, like graduating, getting a job, and moving out of Texas. She says that her

mother describes her as a "seat of the pants" kind of person. Only after the conversation about the Lyon Diet Study did it come up that Kefira's mother has high blood pressure and is on a special diet. When probed about this, Kefira admits that she has a tendency to "tune out" things her mother tells her, for example, why the doctor suggested the diet for her mother. Kefira asserts disbelief in ESP, but not with much conviction, and she laughs throughout the questions preceding her reading of the ESP Study.

Kefira suggests in her written response to her father that grandma should be kept on the medication because, while the results of the study do not give any information about how Makemewell works compared to other medications, it did lower patients' blood pressure. Kefira advises her father that the doctor should know about potential side effects and that he should be asked about those and about the existence of studies that test Makemewell against other blood pressure medications. During the interview, Kefira says that she knows that the medicine is "better than nothing," because the average blood pressure drop of patients on the medication was 19.5 as compared to 9 for patients taking the placebo, so it is doing something. When presented with the results of the test for side effects, Kefira says that the results do not tell her very much because it only provides information about the differences between the medication and nothing and that there are "not really going to be side effects for someone who is just taking a sugar pill." She reiterates the need to have results that test Makemewell against other medications in order to find, for Grandma, the best medication that produces the fewest side effects. Kefira is the only Targeted Study subject to express concern that Makemewell was tested only against a placebo and not against other medications.

Kefira does not specifically mention the anecdotes in her written response. When probed for her reaction to the anecdotes, Kefira is unconcerned by them. She says that she cannot be sure that the medication caused Larry's fever. As far as Sally is concerned,

Kefira says that "everybody reacts differently" and she could be a "random person" for whom the medicine did not work, but it does not necessarily "tell you anything."

Kefira's initial reaction to the Lyon Diet Study article is that it is boring; she seems confused by the vocabulary, knowing that she has heard of "polyunsaturated and monounsaturated" fats and that one is good and the other is bad, but she cannot remember which is which. When asked what she thinks the article was trying to conclude, Kefira says that the researchers are trying to imply that the Mediterranean diet is better, both for heart patients and for healthy people. When asked if the article convinces her of the health benefits of a Mediterranean diet, she says that it does not because of the number of deaths that are recorded. She goes on to say that there is not enough information provided by the study. She would like to know more about the participants because the number of deaths seems very high to her, but if she found out that the participants were all over 70 years old, she would be less concerned about the number of deaths. She reiterates that she is not convinced of the diet's health benefits by the article, but that she believes the diet is healthier because the reduction of intake of red meat and eggs and the increase in intake of fruits and vegetables is what she learned in high school to be important to a healthy diet. When she returns to the numbers in the study, she realizes that she has misread the figures and says that the article is convincing because the number of people on the diet who died or contracted cancer is smaller than the number of people in those groups who were not on the diet.

Kefira's initial reaction to the ESP Study article is that she does not understand the purpose, but she quickly says that the purpose was clearly to find out if there were differences between static and dynamic targets with regard to ESP. She finds it "weird" that there was 37% success rate for the dynamic picture. She is reluctant to believe in ESP, and says that while she was reading the article she was thinking that there must

have been "something they didn't calculate" to obtain the results they did. She considers briefly the fact that there were only 4 target choices as a possible source of the results, but then says that the results mean that she has to think twice about the existence of ESP. When asked what in particular required her to think again about her opinion, she concludes that the low p-values indicate that something other than random guessing was happening in the case of the dynamic targets. She claims that the replications only showed that the procedure was producing such results reliably. In order to find out if the results are due to ESP, she claims that one would need to do different studies to "come at it from a different direction." She is unable to come up with any ideas, although, earlier in the interview she mentions that using four possible targets might be an issue. At the end of the conversation she is not convinced of the existence of ESP, but she is less likely to dismiss it as impossible.

In general, Kefira has correct notions of significance level and knows that small and large p-values mean something about competing hypotheses. Kefira is, however, unable to reliably discuss the meaning of the p-value in context. When asked to explain what the p-value in the Email from Dad task means, she begins by recalling the procedure, speaking softly to herself, "you get a test statistic, and then a p-value, and then a conclusion." Then she says aloud that if you have a "really, really small p-value, then one thing you are testing for is true" and if "you get a high one…the alternative of what you thought was true." She discusses the null hypothesis as the thing that the researchers thought would happen, but then that makes no sense to her because obviously they would have thought that the medicine worked better than the placebo and their results would then indicate that that was not true. When asked if the p-value of .001 was large or small, Kefira says it is very small and that it says the results are significant and that 99.9% of the time, what they expected to happen, that the medication would work better than the

placebo, did indeed happen. Her notion of what the null hypothesis is appears to change based, it seems, on what she believes is probably true. When discussing the Lyon Diet Study, Kefira attends to the raw number of cancers and deaths rather than the p-value. In her discussion of the ESP Study, Kefira refers to the p-values, but discusses them as "statistically significant" without any specificity of meaning in context.

Kefira begins her written response in the Email from Dad task by correctly explaining what "randomized" and "double-blind mean. She does not give any indication of whether or why they are important. The interview is interrupted as Kefira was explaining whether those features were important. When she is reminded that she was talking about whether "randomized and double-blind" were important, she says that the study is not very informative because the medicine was not tested against other medicines. Kefira spontaneously discusses the experimental design of the ESP subject and understands it without further explanation from the interviewer. She says that replication studies only show that something about the set up of the experiment aided in subjects being able to choose the correct dynamic target, but that ESP might not be the underlying cause. Kefira does not discuss experimental design in her comments on the Lyon Diet Study.

References

Alacaci. C. (2004). Inferential Statistics: Understanding expert knowledge and its implications for statistics education. Journal of Statistics Education, 12 (2), downloaded from: http://www.amstat.org/publications/jse/v12n2/alacaci.html

American Heart Association (2005). Lyon Diet Heart Study. Downloaded September, 2005, from http://www.americanheart.org.

Aronson, J. (2002). Stereotype threat: Contending and coping with unnerving expectations. In Improving Academic Achievement: Impact of Psychological Factors on Education, Aronson, J. (ed.). San Diego, CA: Academic Press, pp. 281 – 304.

ARTIST website: http://data.gen.umn.edu/artist//glossary.html

Bem, D. and Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. Psychological Bulletin, 115 (1), pp. 4 – 18.

Ben-Zvi, D. and Garfield, J. (2004). Statistical Literacy, Reasoning and Thinking: Goals, Definitions and Challenges. In The Challenge of Developing Statistical Literacy, Reasoning and Thinking, eds. The Netherlands: Kluwer Academic Publishers, pp. 3 – 16.

Bar-Hillel, M. and Falk, R. (1982). Some teasers concerning conditional probabilities. Cognition, 11, pp. 109 – 122.

Brewer, J.K. (1985). Behavioral statistics textbooks: Sources of myths and misconceptions? Journal of Educational Statistics, 10 (3), pp. 252-268.

Cacioppo, J. and Petty, R. (1982). The need for cognition. Journal of Personality and Social Psychology, 42 (1), pp. 116 – 131.

Cacioppo, J., Petty, R. Feinstein, J. and Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological Bulletin, 119 (2), pp. 197 – 253.

Cacioppo, J., Petty, R. and Kao, C. (1984). The efficient assessment of need for cognition. Journal of Personality Assessment, 48 (3), pp. 306 – 307.

Carpenter, T. (1986). Conceptual knowledge as a foundation for procedural knowledge. In Conceptual and Procedural Knowledge: The Case of Mathematics, James Hiebert, (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 113 – 132.

Carpenter, T. Corbitt, M. and Kepner, H. (1981). What are the chances of your students knowing probability? Mathematics Teacher, 74, pp. 342 – 345.

Chinn, C. and Samarapungavan, A. (2001). Distinguishing between understanding and belief. Theory into Practice, 40 (4), pp. 235 – 241.

Cobb, G. (1992). Teaching Statistics. In *Heeding the Call for Change: Suggestions for Curricular Action*, L.Steen (ed.). Washington, D.C: Mathematical Association of America, , pg 3 – 43.

Cook, D. and Weisberg, S. (1999). Applied Regression Including Computing and Graphics. New York, N.Y.: John Wiley and Sons, Inc.

Cosmides, L. and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition, 58, pp. 1 – 73.

Corsini, R. (1999), The Dictionary of Psychology. USA: Taylor and Francis.

delMas, R. (2004). A comparison of mathematical and statistical reasoning. In The Challenge of Developing Statistical Literacy, Reasoning and Thinking, Ben-Zvi, D. and Garfield, J. (eds.). The Netherlands: Kluwer Academic Publishers, pp. 79 – 95.

delMas, R., Garfield, J. and Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. Journal of Statistics Education, 7 (3). Available online at: http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm

Denes-Raj, V. and Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. Journal of Personality and Social Psychology, 66 (5), pp. 819 – 829.

Dereshiwsky, M. (1999). Electronic Textbook: Limitations and Delimitations. USA: Northern Arizona University, available online at: http://jan.ucc.nau.edu/~mid/edr720/class/methodology/delimitations/

Diseth, A. and Martinsen, O. (2003). Approaches to learning, cognitive style and motives as predictors of academic achievement. Educational Psychology, 23 (2), pp. 195 – 207.

Dweck, C.S. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In Improving Academic Achievement: Impact of Psychological Factors on Education, Aronson, J. (ed.). San Diego, CA: Academic Press, pp. 38 – 61.

Dweck, C.S. (1986). Motivational processes affecting learning. American Psychologist, 41 (10), pp. 1040 – 1048.

Evans, J. St. B. T., Barston, J. L. & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. Memory & Cognition, 11, pp. 295-306.

Evans, J. St B. T., Brooks, P.G., and Pollard, P. (1985). Prior beliefs and statistical inference. British Journal of Psychology, 76, pp. 469 – 477.

Evans, J. St. B. T. and Newstead, S. (1995). Creating a psychology of reasoning: The contribution of Peter Wason. In Perspectives On Thinking and Reasoning: Essays in Honour of Peter Wason. Evans, J. St. B. T. and Newstead, S., (eds.). East Sussex, UK: Lawrence Erlbaum Associates, Ltd.

Evans, J. St. B. T., Venn, S., and Feeney, A. (2002). Implicit and explicit processes in a hypothesis testing task. British Journal of Psychology, 93, pp. 31 – 46.

Everitt, B. (1993). Cluster Analysis, Third Edition. Cambridge, UK: Cambridge University Press.

Fischhoff, B. (1982). Debiasing. In Judgment Under Uncertainty: Heuristics and Biases, Kahneman, D., Slovic, P. and Tversky, A. (eds.). Cambridge, UK: Cambridge University Press, pp. 422 – 444.

Franko-Watkins, A., Derks, P. and Dougherty, M. (2003). Reasoning in the Monty Hall problem: Examining choice behavior and probability judgments. Thinking and Reasoning, 9, (1), pp. 67 – 90.

GAISE (2006). College Report of the Guidelines for Assessment and Instruction in Statistics Education Project downloaded from: http://www.amstat.org/education/gaise/GAISECollege.htm

Garfield, J. (2003). Assessing statistical reasoning. Statistics Education Research Journal, 2 (1), pp. 22 – 38. Available online: http://fehps.une.au/serj

Garfield, J. and Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for Research. Journal of Research in Mathematics Education, 19 (1), pp. 44 – 63.

Garfield, J. and Gal, I. (1999). Assessment and statistics education: Current challenges and directions. International Statistical Review, 67 (1), pp. 1 – 12.

Girotto, V. and Gonzalez, M. (2001). Solving probabilistic and statistical problems: a mater of information structure and question form. Cognition, 78, pp. 247 – 276.

Hertwig, R. and Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. Journal of Behavioral Decision Making, 12 (4), pp. 275-305.

Hiebert, J. and Lefevre, P. (1986). Conceptual and Procedural Knowledge in Mathematics: An Introductory Analysis. in Conceptual and Procedural Knowledge: The Case of Mathematics, James Hiebert, (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 1 – 27.

Hiebert, J. and Wearne, D. (1986). Procedures over concepts: The acquisition of decimal number knowledge. In Conceptual and Procedural Knowledge: The Case of Mathematics, James Hiebert, (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 199 – 224.

Hirsch L. and O'Donnell A. (2001). Representativeness in statistical reasoning: Identifying and assessing misconceptions. Journal of Statistics Education, 9 (2). Available online at: http://www.amstat.org/publications/jse/v9n2/hirsch.html.

Jordan, J. (2004). The use of writing as both a learning and an assessment tool. Presentation at ARTIST Roundtable Conference. Lawrence University, Appleton, WI. Available online at: http://www.rossmanchance.com/artist/Proctoc.html

Johnson, R. and Wichern, D. (1998). Applied Multivariate Statistical Analysis, Fourth Edition. Upper Saddle River, N.J.: Prentice-Hall, Inc.

Johnson-Laird, P. (1994). Mental models and probabilistic thinking. Cognition, 50, pp. 189 – 209.

Kahneman, D. and Tversky, A. (1982). On the study of statistical intuitions. Cognition, 11, pp 123 – 141.

Kaplan, J. (2001). Regression and Cluster Analysis of Demographic Data: A Study of Clustering Techniques Used in Creating Campus Comparison Groups. Unpublished Master's Report. Austin, TX: The University of Texas at Austin.

Koehler, J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. Organizational Behavior and Human Decision Processes, 56, pp. 28 – 55.

Krettenauer, T. (2005). The role of epistemic cognition in adolescent identity formation: Further evidence. Journal of Youth and Adolescence, 54 (3), pp. 185 – 198.

Kuhberger, A. (1998). The influence of framing effects on risky decisions: A meta-analysis. Organizational Behavior and Human Decision Processes, 75 (1), pp. 23 – 55.

Kuhn, D. (2001). How do people know? Psychological Science, 12 (1), pp. 1 – 8.

Kuhn, D., Cheney, R. and Weinstock, M. (2000). The development of epistemological understanding. Cognitive Development, 15, pp. 309 – 328.

Lane-Getaz, S. (2005). Reasoning about P-values: Pilot survey. Unpublished Manuscript. Available online: www.causeweb.org/uscots/spotlight/files/R6.pdf

LeBoeuf, R. and Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility of framing effects. Journal of Behavioral Decision Making, 16, pp. 77 – 92.

Martin, B., Sherrard, M. and Wentzel, D. (2005). The role of sensation seeking and need for cognition on web-site evaluations: A resource matching perspective. Psychology and Marketing, 22 (2), pp. 109 – 126.

Mason, L. and Boscolo, P. (2004). Role of epistemological understanding and interest in interpreting a controversy and in topic specific belief change. Contemporary Educational Psychology, 29, pp. 103 – 128.

Meyer, R. and Wittrock, M. (1996). Problem-solving transfer. In D.C. Berliner & R.C. Calfee (eds.), Handbook of Educational Psychology. New York, N.Y.: MacMillan, pp. 47 – 62.

Miller, P. and Fagley, N. S. (1991) The effects on framing, problem variations, and providing rationale on choice. Personality and Social Psychology Bulletin, 17 (5), pp. 517 – 522.

Mills, J. (2002). Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. Journal of Statistics Education, 10 (1), available online: http://www.amstat.org/publications/jse/v10n1/mills.html

Minstrell, J. and Krauss, P. (2005). Guided inquiry in the science classroom. . In How Students Learn, Bransford, J. and Donovan, S. (eds.). Washington, D.C.: National Academy Press, pp. 475 – 514.

Montgomery, D. (1997). Design and Analysis of Experiments, Fourth Edition. New York, N.Y.: John Wiley and Sons, Inc.

Moore, D. (2003). The Basic Practice of Statistics, Third Edition. New York, NY: W.H. Freeman and Company.

Moore, D. (1990). Uncertainty. In One the Shoulders of Giants: New Approaches to Numeracy, Steen, L., (ed). Washington, D.C.: National Academy Press, pp. 95 – 137.

Murray, T., Clermont, Y. and Binkley, M. (2005). Measuring Adult Literacy and Life Skills: New Frameworks for Assessment. Ottawa, Canada: Minister of Industry. Available online: http://www.nald.ca/fulltext/measlit/cover.htm

Mynatt, C. Doherty, M. and Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. The Quarterly Journal of Experimental Psychology, 46A (4), pg. 759 – 778.

National Research Council, (2003). Learning and Instruction: A SERP Research Agenda. Donovan, M. S. and Pellegrino, J. W., (eds.). Washington, D.C.: National Academy Press, Available online at: http://www.nap.edu/openbook/0309090814/html/index.html

National Research Council, (2001). Adding + It Up, Jeremy Kilpatrick, Jane Swafford, and Bradford Findell, (eds.). Washington, D.C.: National Academy Press, Available online at: http://www.nap.edu/openbook/0309069955/html/index.html

National Research Council, (1996). National Science Education Standards, National Committee on Science Education Standards and Assessment. Washington, D.C.: National Academy Press, Available online at: http://books.nap.edu/catalog/4962.html

National Research Council, (1999). How People Learn: Brain, Mind, Experience, and School, John Bransford, Ann Brown, and Rodney Cocking, (eds.). Washington, D.C.: National Academy Press. Available online at: http://www.nap.edu/

Nunnally, J. (1967). Psychometric Theory. New York, N.Y.: McGraw-Hill, Inc.

Peracchio, L. and Meyers-Levy, J. (2005). Using stylistic properties of ad pictures to communicate with consumers. Journal of Consumer Research, 32, pp. 29 – 40.

Ramsey, F. and Schafer, D. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis, second edition. U.S.A.: Thomson Learning, Inc.

Resnick, L. B. (1987). Education and Learning to Think. Washington, D. C.: National Academy Press, Available online: http://www.nap.edu/books/0309037859/index.html

Rhoads, T. and Murphy T.J., (2004). Statistics Concept Inventory. Paper presented at ARTIST Roundtable Conference on Assessment in Statistics, Appleton, WI. Available online at: http://www.rossmanchance.com/artist/Proctoc.html

Sadowski, C. and Gulgoz, S. (1992). Internal consistency and test-retest reliability of the need for cognition scale. Perceptual and Motor Skills, 74 (2), pp. 610.

Schoenfeld, A.H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. Grouws (ed.), Handbook of Research of Mathematics Teaching and Learning. New York, N.Y.: MacMillan, pp. 334 – 370.

Sedlmeier, P. (1999) Improving Statistical Reasoning : Theoretical Models and Practical Implications. Mahwah, N..J.: Lawrence Erlbaum Associates, Inc.

Silver, E. (1986). Using conceptual and procedural knowledge: A focus on relationship. In Conceptual and Procedural Knowledge: The Case of Mathematics, James Hiebert, (ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates, pg 181 – 198.

Sloman, S. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119 (1), pp. 3 – 22.

Stanovich, K. (1999). Who Is Rational?: Studies of Individual Differences in Reasoning. Mahwah, N..J.: Lawrence Erlbaum Associates, Inc.

Stanovich, K. and West, R. (1998a). Individual differences in framing and conjunction effects. Thinking and Reasoning, 4 (4), pp. 289 – 317.

Stanovich, K. and West, R. (1998b). Individual differences in rational thought. Journal of Experimental Psychology: General, 127 (2), pp. 161 – 188.

Stanovich, K. and West, R. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. Journal of Educational Psychology, 89 (2), pp. 342 – 357.

Strauss, A. and Corbin, J. (1998) Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: Sage Publications.

Tubau, E. and Alonso, D. (2003). Overcoming illusory inferences in a probabilistic counterintuitive problem: The role of explicit representations. Memory and Cognition, 31, (4), pp. 596 – 607.

Thompson, V. (1996). Reasoning from false premises: The role of soundness in making logical deductions. Canadian Journal of Experimental Psychology, 50 (3), pp. 315 – 319.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90 (4), pp. 293 – 314.

Wallman, K. (1993). Enhancing statistical literacy: Enriching our society. Journal of the American Statistical Association, 88 (421), pp. 1 – 8.

Weinstock, M. (2005). Cognitive bases for effective participation in democratic institutions: Argument skill and juror reasoning. Theory and Research in Social Education, 33 (1), pp. 73 – 102.

Weinstock, M. and Cronin, M. (2003). The everyday production of knowledge: Individual differences in epistemological understanding and juror-reasoning skill. Applied Cognitive Psychology, 17, pp. 161 – 181.

Weisstein, E. (2006) "Bonferroni Correction." From MathWorld--A Wolfram Web Resource. Available online at: http://mathworld.wolfram.com/BonferroniCorrection.html

Wild, C.J. and Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67(3), pp. 223 – 265.

Vita

Jennifer Julia Kaplan was born in Landstuhl, Germany on September 28, 1967, the daughter of Ellen B. Kaplan and Michael S. Kaplan. She graduated from The MacDuffie School, Springfield, MA, in 1985 and from Brandeis University, with a B.A. in mathematics and a high school teaching certificate, in 1989. Jennifer taught high school mathematics for ten years, the first three in public schools systems in the Boston, MA suburbs. In 1992, Jennifer moved to Istanbul, Turkey to teach at the Istanbul International Community School. In 1994, she relocated again, this time to Caracas, Venezuela, where, in 1998 she was appointed the head of the mathematics department at Colegio Internacional de Caracas. In 1999, she returned to the U.S. and began graduate studies in the Department of Mathematics at The University of Texas at Austin. After earning an M.A. from that department, she remained in Austin and in the Department of Mathematics to complete her Ph.D. During most of her tenure at UT-Austin, Jennifer was an Assistant Instructor in the Department of Mathematics, teaching the mathematics content course for elementary education majors. Jennifer has accepted a position as an assistant professor in the Department of Statistics and Probability and the Division of Science and Mathematics Education at Michigan State University to start in August 2006.

Permanent address: 128 South Water Street, Edgartown, MA 02539

This dissertation was typed by the author.